COMPARING RANDOM FOREST WITH GENERALIZED LINEAR REGRESSION: PREDICTING CONFLICT EVENTS IN WESTERN AFRICA

by

Tyler M. Gill

A Thesis Submitted to the Faculty of the

SCHOOL OF GOVERNMENT & PUBLIC POLICY

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF ARTS

In the Graduate College

THE UNIVERSITY OF ARIZONA

2019

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Master's Committee, we certify that we have read the thesis prepared by Tyler Gill, titled *Comparing Random Forest with Generalized Linear Regression: Predicting Conflict Events in Western Africa* and recommend that it be accepted as fulfilling the thesis requirement for the Master's Degree.

Kirssa Cline Ryckman — Date: 12/18/2019

Paulette Kurzer — Date: 12/18/2019

Jessica Maves Braithwaite — Date: 12/18/19

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to the Graduate College.

We hereby certify that we have read this thesis prepared under our direction and recommend that it be accepted as fulfilling the Master's requirement.

Kirssa Cline Ryckman — Date: 12/18/19
Master's Thesis Committee Co-Chair
School of Government and Public Policy

Paulette Kurzer — Date: 12/18/19
Master's Thesis Committee Co-Chair
School of Government and Public Policy

Table of Contents

List of Figures and Maps

Abstract

Despite the progress of conflict prediction models within the last two decades, most approaches lack a spatial approach to conflict prediction. This study explores the intersection of geography and conflict prediction, while discussing how spatial analysis and conflict prediction can integrate. The objective is to compare a spatial machine learning method, Classification Random Forest, to a more traditional statistical method, Logistic Generalized Linear Regression, to assess the sub-national predictive power of conflict occurrence in Western Africa. The two models are fitted to subnational data for Western Africa at the district level covering the years 2015 – 2017, generating an out of sample prediction for 2018. The overall accuracy is assessed using an *F*-1 score, which accounts for sensitivity and precision of the models, to discover areas of over-prediction and under-prediction. The Random Forest model produced an *F*-1 score of 0.58582, while the Generalized Linear Regression model had an *F*-1 score of 0.61017. A significant difference between the two models was not detected. The Generalized Linear Regression had a better overall accuracy, but the Random Forest model predicted more incidences of conflict correctly. Between the two models, five explanatory variables contributed to the predictive power of both models: Conflict Density, Road Density, Area of the District, Nighttime Lights, and Population Density. Future research should also explore the effect of conflict hot spots on conflict prediction models. Given the success of the Generalized Linear Regression model, the next logical step is to explore the local variation with a Geographic Weighted Regression model.

Introduction

The role of geography is understudied in quantitative international relations literature, especially in the subfield of conflict prediction and forecast. Spatial analysis and conflict & machine learning rarely integrate, but the combination of spatial analysis and machine learning methods for conflict prediction is an untapped area of research that could yield powerful prediction results. Civil conflict scholars have mixed opinions on the effects of geographic features on conflict; a spatial approach toward conflict predictions could uncover the importance of common geographic and socio-economic variables in association with the occurrence of civil conflict. The variables associated with conflict and approaches employed in the previous literature are often mismatched when testing for the geographic aspect of conflict. In well-cited civil conflict literature (Fearon & Laitin, 2003; Collier & Hoeffler, 2004), the geographic variables are usually proxies at the state level, such as percent of mountainous terrain for a state, but rarely do studies model the actual terrain values. Furthermore, there remains subnational variation of conflict within states. Subnational spatial analysis has provided some answers into how conflict interacts with space, but it is rarely incorporated or tested for predictions.

Study of the intersection of geography and international relations dates back to the idea of geographic determinism and the formation of the sub-discipline geopolitics (Kjellen, 1916). Diehl (1991) provides an in-depth empirical examination of geography and war. Three concepts are essential to the theoretical background of geography and interstate conflict: proximity (Bremer, 1992; Starr, 1978; Richardson, 1960), territory (Diehl, 1991; Prescott, 1965; Mandel, 1980; Huth, 1996), and distance (Gleditsch, 1995; Boulding, 1962; Bueno de Mesquita, 1981; Gleditsch & Singer, 1975).

Studies of geography and intrastate conflicts differ greatly from studies of geography and interstate conflict. Most of the empirical theoretical background of civil wars is based on the theories of greed and grievance (Collier & Hoeffler, 2004) and opportunity (Fearon & Laitin, 2003). While the scale of analysis of intrastate conflict is within a state, the literature places a focus on physical geography characteristics, such as rough terrain (Buhaug, 2005; DeRouen & Sobek, 2004) and natural resources (Auty, 2004; Collier & Hoeffler, 2004; Berdal & Malone, 2000; Addison et al., 2002) and human geography characteristics, such as identity (Denny & Walter, 2014; Weidmann, 2009; Hegre et al., 2001; Elbadawi & Sambanis, 2002; Ellingsen, 2000) & population size (Buhaug, 2005; Sambanis, 2004; Herbst, 2000). The existing literature on intrastate conflicts provides the justification for the variables used in this study. The incorporation of distance into intrastate conflict literature is recent (Buhaug & Gates, 2002; Raleigh & Hegre, 2009) but important because the concept is used across the different levels of conflict (interstate, state, and sub-national) and vital to spatial analysis of conflict.

Spatial analysis of international conflict took off in the late 1980s, utilizing the approach of spatial econometrics (Paelinck & Klaassen, 1979; Anselin, 1988), which explores spatial interaction (spatial autocorrelation) and spatial structure (spatial heterogeneity) within the data, typically using regression models (O'Loughlin & Anselin, 1991; Kirby & Ward, 1987; Cliff & Ord, 1973; 1981). Local Indicators of Spatial Analysis (LISAs) were later developed to identify clustering of events (Anselin, 1995; Ord & Getis, 1995). The literature establishes that conflict on the international scale clusters in both space and time (Houweling & Siccama, 1985; 1988; Braithwaite, 2010a) and in civil violence (Buhaug & Gleditsch, 2008; Khatiwada, 2014). The evolution of spatial econometrics helped grow the concept of diffusion; it expanded from Most & Starr's (1980) framework of "opportunity and willingness." Diffusion and international conflict

studies (Starr & Most, 1983; 1985; Faber, Houweling & Siccama, 1984) & diffusion and intrastate conflict studies (Raleigh, Witmer, & O'Loughlin, 2010; Schutte & Weidmann, 2011; Murdoch & Sandler, 2002; Salehyan & Gleditsch, 2006) provides the last theoretical concept for a spatial approach to conflict prediction.

Geographic Information Systems (GIS) has allowed researchers the ability to analyze spatial relationships in international relations in a more advanced fashion than that of traditional studies. The advances of GIS have placed studies into two broad categories: studies that focus on the locations of conflicts (Braithwaite, 2010a; Gleditsch et al., 2002; Sundberg & Melander, 2013) and studies that focus on the violence within conflicts (De Juan, 2012; Flint et al., 2009; Theisen et al., 2012; Buhaug & Rod, 2006; Hegre & Raleigh, 2009; Buhaug, Gates, & Lujala, 2009; Lujala, 2009; Murshed & Gates, 2005; Do & Iyer, 2010).

Machine learning and conflict literature predates spatial analysis and conflict, yet there has still been little interaction between the two disciplines. The first generation of conflict prediction is characterized by the efforts to collect scientific data on conflict to produce early-warning systems (Wright, 1942; Sorokin, 1957; Richardson, 1960), while the second generation is characterized by the ability to build statistical models based on news source data (Schrodt, 1988; Schrodt, Davis & Weddle, 1994; Schrodt, 1991a; O'Brien, 2010). The third generation grew into the conflict prediction literature of today utilizing different models and techniques, such as Random Forest (Hill & Jones, 2014; Blair & Sambanis, 2016; Muchlinki, Siroky, He, & Kocher, 2015; Perry, 2013; Lazicky, 2017; Blair, Blattman & Hartman, 2017), logistic regression models (Ward & Gleditsch, 2002; Weidmann & Ward, 2010; Weezel, 2017; Celiku & Kraay, 2017), neural networks (Schrodt, 1991b; Beck, King, & Zeng, 2000), time series models (Goldstein, 1992; Schrodt & Gerner, 2000; Brandt & Freeman, 2005; Brandt, Freeman, &

Schrodt, 2014; Yonamine, 2013), Hidden Markov Models (Schrodt, 1999; Schrodt, 2000; Shearer, 2007; Schrodt, 2006; Schrodt et al., 2013), point processes (Schutte, 2016), and climate-sensitive models (Witmer et al., 2017).

Previous literature on conflict predictions has been criticized for a lack of theory driven work (Jones & Linder, 2015; Blattman & Miguel, 2010; Basuchoudhary et al., 2018). The incorporation of a geospatial approach provides the theoretical framework for spatial conflict prediction. Proximity, territory, distance, and diffusion are the basic concepts of the intersection between geography and international conflict. The concepts can be incorporated into conflict prediction at the subnational level. Most of the previous research into conflict predictions and forecast explores violence or conflict outbreak temporally, while spatial econometric approaches in conflict studies focuses on identifying clusters, or hot spots, of conflict within different regions; the next step is to explore where conflict is going to occur based on past incidents. Once future conflict locations are predicted, it is important to easily relay the information to policymakers for action. Visualizing the spatial conflict predictions using cartography or maps, can easily inform policymakers of areas that require further attention.

Quantitative geospatial analysis at the subnational level has expanded recently (Buhaug & Gates, 2002; Kalyvas, 2006; Raleigh et al., 2010) and to a lesser extent, so has conflict prediction at the subnational level (Weezel, 2017; Schutte, 2016). Despite gains in both spatial analysis and conflict and machine learning and conflict, why has the literature so rarely integrated the approaches of both? Both have been in silos with little cross-interaction, the result being that literature on spatial conflict prediction is in its infant stage. Predictive models that take geographic considerations into account are few in the civil war literature (Raleigh et al., 2010). Subnational geospatial analysis typically examines conflict incidences and intensity, as opposed

to onset, and is divided into three types of studies: analysis on the locations of conflicts, analysis on the location of violence within conflicts, and analysis of the diffusion of violence (De Juan, 2012). Conflict prediction literature has just begun to incorporate geography or spatial disaggregation into the modelling approaches (Perry, 2013; Weezel, 2017; Weidmann & Ward, 2010; Schutte, 2014). The advances of GIS and machine learning models has allowed researchers the ability to analyze conflict occurrence using methods previously unavailable.

This study objectives are to compare two spatial conflict prediction models based on a Classification Random Forest model and a Logistic Generalized Linear Regression model and assess which model produces superior predictive power. In doing so, the spatial explanatory variables are evaluated for their explanatory power in the model creation. Based on past performance of Random Forest models and Logistic Regression methods in the previous literature, it is expected that the Random Forest model will outperform the Generalized Linear Regression Model.

The spatial Classification Random Forest model and Logistic Generalized Linear Regression model utilize Environmental Systems Research Institute (ESRI) software, ArcMap. The dependent variable is political violence events from 2015 – 2017 from the Armed Conflict Location and Event Data Project (ACLED; Raleigh et al., 2010), aggregated into administrative two boundaries (districts) in Western Africa. The explanatory variables are divided into three groups: explanatory variables, explanatory distance variables, and explanatory rasters. The explanatory variables include: ethnic composition, ethnic fractionalization, language composition, presence of diamonds, presence of Petroleum, presence of previous Conflict vents from 2012 – 2015, presence of foreign terrorist organizations, polity IV score, and area of district. The explanatory distance variables are: distance to the capital city, distance to the

border, distance to the natural resources of diamonds and distance to the natural resources of petroleum. The explanatory rasters comprise of: land cover, night light emissions, elevation data, slope, population density, infant mortality rate, percent of children aged 0 - 14, conflict density, road density and accessibility of road infrastructure.

The Classification Random Forest model and Logistic Generalized Linear Regression model are trained using an in-sample technique of political violent event incidence locations aggregated to administrative two boundaries from 2015 – 2017, with 30% of the data excluded for validation. Once the models are trained, the predictive capabilities are tested utilizing an out of sample technique against the 2018 ACLED data. A confusion matrix is produced and the *F*-1 scores are calculated to assess the overall prediction accuracy; both results are also mapped to see the spatial variation. The Classification Random Forest model produces a Top Variable Importance list and the Logistic Generalized Linear Regression signifies statistically significant variables; the important variables from both models are then compared.

Surprisingly, the Logistic Generalized Linear Regression model produced a better *F*-1 score than the Classification Random Forest model. The Logistic Generalized Linear Regression model had a *F*-1 score of 0.61017 and the Classification Random Forest model had a *F*-1 score of 0.58582. The Random Forest predictive power is similar to Perry (2013) Random Forest model, but the Generalized Linear Regression *F*-1 score is significantly better than a similar study, Weezel (2017), which produced a *F*-1 score of 0.257. Five explanatory variables were significant to both models: Conflict Density, Road Density, Population Density, Nighttime Lights, and Area of the District. Overall, the two models had similar predictive power, with the Random Forest predicting more true conflict occurrence districts but the Generalized Linear Regression predicting more true non-conflict occurrence districts. The Generalized Linear

Regression model is a global model, future research could build upon the model and include a local model, Logistic Geographic Weighted Regression.

The literature review follows the introduction; it is divided into four sections. First, an overview of the theoretical foundation based in the literature of geography and interstate conflict. Second is the literature of geography and intrastate conflict. Followed by an overview of spatial analysis of conflict, including at the subnational level. Finally, a review of the conflict prediction and forecasting in quantitative international relations is included. The methodology section follows the literature review, with an overview of the underlying statistics in the Random Forest model and Generalized Linear Regression model, an overview of the explanatory variables and a summary of the procedures. The results and discussion of the results follow the methodology section. A conclusion identifies areas of future research.

Chapter One: Geography in Interstate Conflict Studies

This section provides a background of the nexus between geography and international relations at the state, dyad, and system level. Early literature on geography and conflict focused on interstate conflict, mainly in the form of wars. When scholars first began to consider geography in the study of international relations, geopolitics was born with an emphasis on geographic determinism. Soon after, scholars began to evolve from a determinist position toward geography to one that viewed geography as a facilitator of conflict. The concepts of territory, contiguity or proximity, and distance are three geographic concepts central to the study of geography and interstate conflict. There are two groups of arguments in the literature on the relationship between international politics and geography. The first group of literature focuses on proximity, with studies on contiguity. The second group of literature concentrates on opportunity as a result of proximity. All three geographic concepts are interrelated, yet each contribute individually to the theoretical foundation of spatial analysis.

1.1 Geopolitics

Early studies of geography and international relations focus on the idea of geographic determinism. Geography's first interaction with international relations was focused on the state; early geopolitical thought argued geography was a primary determinant of state behavior. The term geopolitics was first coined by Rudolf Kjellen in 1899 and described as "the theory of a state as a geographical organism or phenomenon in space" (Kjellen, 1916). According to Cohan (2015), geopolitics as a vehicle for integrating geography and international politics finds it useful to define geopolitics as not a school of thought, but as a mode of analysis, relating diversity in context and scale of geographic settings to identify spatial framework in which power flows. It is important to note the early geopolitical thought as a foundation for the theoretical frameworks of

the twentieth century; this study utilizes geopolitics as mode of analysis and a foundation to the theoretical framework.

There are a number of scholars that influenced the study of geography and war; the four most influential theorists of geopolitical thought were Mahan, Mackinder, Spykman, and Haushofer (Diehl, 1991; Cohen, 2015) These scholars created the foundation of modern geopolitics that built the theory behind geography and international relations. Admiral Alfred T. Mahan was one of the first American scholars to incorporate geography into international relations. He focused mainly on sea power and argued the critical zone of conflict lay where Russian land power and British sea power met between the thirtieth and fortieth parallels in Asia. Halford Mackinder was a British geographer who created one of the first geopolitical theories. Mackinder (1904) created the famous geopolitical theory describing the "pivot area" of the world, based on geographic realities. "Whoever rules East Europe, will rule Heartland. Whoever rules the Heartland, will rule the World Island. Whoever rules the World Island, will rule the world." The heartland theory was among the first theories to incorporate geography into international relations. Both theories from Mahan and Mackinder combined with the German defeat of World War I formed German *geopolitik* led by Karl Haushofer. *Geopolitik* was based on geographic determinism of the state. As *geopolitik* merged into the ideological foundations of the Nazi regime, geopolitics as a discipline began to lose credibility. After World War II, scholars began to shift from geographic determinism toward geography as a factor, which provides the foundation for modern studies of geography and international conflict.

Early literature on geography and conflict focused on international conflict, mainly in the form of wars. Geography was seen as a factor in the complex international system, in which states competed, causing international conflict. For many scholars, power is a central element to

understanding international relations. Morgenthau (1967) identifies geography as the first factor for national power. While Morgenthau (1967) focused on physical geography, such as natural barriers like rivers, mountains, and oceans, natural barriers are used as independent variables in later research. According to Nye Jr. (2003), in an international system, the distribution of power among states helps make predictions about certain aspects of a states' behavior.

Toward the middle of Twentieth Century scholars began to shift from geographic determinism. Scholars began to recognize that international behavior is complex and requires more than the absolutist position of geographic determinism. Sprout & Sprout (1965) coined the term 'environmental possibilism,' with a focus on how geography does not determine certain actions a state will make. According to Diehl (1991), the states respond to their environment, but the environment does not condition or compel policymakers to perform particular actions. Sprouts work provides the initial framework of the theory that geography does not dictate wars that occur between two states. Later studies built on this framework to set the environment for spatial analysis of conflict.

1.2 Proximity, Territory, Distance

The three geographic concepts of proximity, territory and distance have been central fixtures in studies of geography and international conflict. Contiguity and territory both have considerable empirical literature on the concepts yet have been treated as independent concepts of each other (Senese, 2005). Distance is the third geographic concept of international conflict; this has less literature dedicated to it than contiguity and territory but has the same importance to the foundation of modern geographic studies of conflict.

Contiguity or proximity is an important factor when examining the link between geography and international conflict. Contiguity is based theoretically on Tobler's (1970) first

law of geography, “everything is related to everything else, but near things are more related than distant things.” The first law of geography provides the foundation for contiguity or proximity studies. Bremer (1992) recognized contiguity as the most important factor for creating dangerous dyads between two states. Building upon the work of ‘environmental possibilism’, Starr (1978) created the framework of “opportunity” and “willingness”. Opportunity is the possibility of interaction between entities, such as states and willingness is the processes that lead to the opportunities to go to war. Opportunity is most closely related to the geographic concept of proximity. A higher level of interaction between states leads to a higher risk of conflictual interactions. The “opportunity” and “willingness” framework provided the foundation for contiguity studies examining the borders between states and how it relates to international conflict. This is important because under this framework, empirical studies of borders began to model conflict diffusion and patterns between states. These studies provided the foundation of spatial autocorrelation, later discussed in section three.

Closely related to the concept of proximity is the concept of borders. The number of shared borders a state possesses is one variable thought to determine geographic opportunity. Richardson (1960) found the more borders a state had, the more likely it was to experience war and conflict; whereas Midlarsky (1975) found the greater number of borders a state possesses, the more uncertainty the state faces. The greater number of borders increases the interactions between states, which in turn increases the chance of conflict. Starr & Most (1976; 1983; 1985) expand on the framework on interactions but suggest borders themselves do not cause conflicts, but rather structure the risks and opportunities for conflictual behavior to occur. The opportunity framework is important when examining conflict at the international scale, but it is largely descriptive; related to the concept of contiguity is territory.

Territory is the oldest concept with roots to geopolitical thought of geographical determinism; during this time, scholars focused on the quest of great powers to acquire strategic territory around the globe (Diehl, 1991). Spykman (1938) analyzed geography based on the size of states and the locations of states. "Geography is the most fundamentally conditioning factor in the formulation of national policy because it is the most permanent (Spykman, 1938)." Furthermore, Starr (2005) describes two purposes territory serves in the study of international relations. First, by defining the territorial extent of political units, territory indicates the physical – geographic distance between the units. Second, territory provides an important component of "group identity". Historically, geography was synonymous with territory; many studies look at geography as a source of conflict, where states fight over specific geographic areas. Diehl (1991) describes how studies that focus on geography as a source of conflict have been concerned with the characteristics or conditions that surround territorial disputes. Prescott (1965) examined the change in relative power of states as an important predictor of territorial conflicts, while Mandel (1980) focused on the characteristics of the states involved, such as the frequency of border disputes associated with an equal distribution of power between disputants and low technology states. Both studies examined the relative power of the states in territorial conflicts.

The characteristics of the territory are important components in affecting the willingness of states to fight over territory (Diehl, 1991). There remains some debate whether ethnicity or resources drive territorial disputes. Mandel (1980) claims the severity is driven by ethnic concerns, while Koch, North, & Zinnes (1960) note that conflict over resources can occur if there are great values of the resources in the territory and a dispute over the allocation of those resources. Furthermore, Nordquist (1986) determined ethnic factors do not play a greater role in border conflicts compared to non-border conflicts. Huth (1996) found that states are in territorial

disputes because domestic political concerns drive foreign leaders to adopt policies that risk conflict; the finding placed importance on contiguity or proximity of states.

Although there is no consensus on why states fight over territory, territorial issues still remain an important source of conflict. The concept of territory provides a theoretical framework for research questions about the locations of conflict, such as why conflicts are clustered in a particular location and not in other locations and the characteristics of the conflicts, which is beneficial for modeling the conflicts. The concept of territory has evolved from the importance of a state holding territory to the concept of territorial disputes between states. Despite the evolution of the term, the number of studies on territorial disputes is rather limited within the study of conflict.

Distance is a factor related to territory and contiguity and is important for two reasons: distance is the key to creating neighborhoods for spatial analysis and distances of objects are used as independent variables within the analysis for conflict locations. The examination of distance as a function of geography within international relations studies is based on the gravity model of trade (Isard, 1954). The gravity model of trade is a spatial interactional model that formalizes interaction between two areas as a function of distance and expected propensity to interact. Gleditsch (1995) describes the theoretical justification for the relationship between distance and interaction based on cost, time, and intervening as opposed to state border length.

The gravity model of trade evolved into a form of power projection between states, thanks in large part to Boulding's (1962) proposed framework of Loss of Strength Gradient (LSG), which demonstrated the importance of geographic distance by showing that the farther away the target of aggression, the less strength could be made available. Furthermore, proximate states will be perceived as more threatening than those that are far away because distant states

are less visible (Boulding, 1962). The relevance of LSG in the modern international system is debated but Webb (2007) argues distance is still important because of the competitive nature of war and the impermanence of great-power status. LSG is also the foundation behind the use of inverse distance weighting, a method of conceptualizing the spatial relationship in spatial cluster analysis.

Bueno de Mesquita (1981) uses the framework of LSG to develop the expected-utility theory, which is used to test neighboring conflicts from long distance conflict; he finds long-distance wars are significantly initiated by great powers. Early studies of distance focused power projection in interstate war. This is sometimes measured by distance of capitals between states. Gleditsch & Singer (1975) found that the average distance between the capitals of warring dyads is considerably shorter than the average distance between all pairs of states in the system. The findings on average distance in interstate conflict literature is also applied to intrastate conflict and sub-national conflict literature.

Chapter Two: Geography in Intrastate Conflict Studies

This section will discuss the role of geography in civil war literature. Studies of geography and intrastate conflicts vary greatly from studies of geography and interstate conflict, yet the theoretical rationale remains similar. Studies of intrastate conflict focus on physical geography characteristics, such as rough terrain & natural resources and human geography characteristics such as population size & ethnicity, whereas studies on interstate conflict focus on territory, proximity, and distance. The theoretical framework for empirical studies on civil war and geography is often motivated by how local factors such as rebels' access to easily exploitable natural resources and sanctuaries in rough terrain increase the viability of insurgency (Buhaug & Lujala, 2005). Distance remains important for both interstate conflict studies and civil war studies. Intrastate conflict studies provide an important theoretical framework for conflict prediction.

Buhaug (2005) developed a typology of the literature of intrastate conflict. The literature of intrastate conflict, or civil wars, is shaped by three factors: motivation, opportunity, and identity (Gurr 1970; Ellingsen, 2000; Collier & Hoeffler, 2004). The three factors have a mutual dependence. Motivation builds upon Starr's (1978) willingness concept. Rebellion is explained not by absolute levels of poverty or inequality but the perceived unfavorable conditions of a particular group. Economic theories of civil war emphasize 'greed' or personal ambition (Collier & Hoeffler, 2004). Opportunity refers to the degree to which rebellion is a feasible means to redress the grievance; rebellions have to have opportunity to challenge the government. Opportunity, as it relates to financing a rebellion has received the most attention (Collier & Hoeffler, 2004). The most common sources of finance include extraction of natural resources, donations from diasporas, and subventions from foreign governments. Rough terrain is

considered an opportunity factor because it could offer vital protection from government forces and access to foreign soil, which in addition to providing safe havens facilitates trade and smuggling. Common identity maintains the coherence of the rebel group. A good amount of civil conflict literature focuses on ethnic groups and ethnic conflict. Identity conflicts could start as intercommunal violence but can also be the product of actions by a government against ethnic minorities. A group's identity does not need a long history but rather a common desire (economical, political, or ideological) to create change of the dominant group.

2.1 Physical Geography

Geographic features in civil war studies have focused on physical geography characteristics, such as natural resources and terrain. Both natural resources and terrain have theoretical augments linking them to civil wars and are used as variables in quantitative studies. Typically, in the civil war literature, physical geographic features are modeled using proxies aggregated to the state level. This study will explore the subnational variation of the physical geographic features.

Natural Resources

Natural resources are commonly linked to civil war onsets, yet the empirical literature shows mixed results. Buhaug (2005) claims the link between natural resources and risk of armed conflict is the most controversial part of the geography-civil war nexus. There are two types of focuses for the link between natural resources and conflict. The first group claims a causal relationship between resource scarcity and conflict; this group focuses on renewable resources, such as water and soil and is occupied with future scenarios. The second group, used more in the literature, explores the relationship between valuable, non-renewable resources, such as fuels, gems, and drugs.

Government control over natural resources varies greatly by the type. There are point resources, such as oil drilling operations and pit mining of minerals; these are non-renewable and geographically concentrated (Buhanug & Gates, 2002; Addison et al., 2002). The second type of resources are diffuse resources, such as timber resources, illegal drugs, and soils & water. These are renewable and geographically spread. Diffuse resources are more widely dispersed and more difficult for a government to control. De Soysa (2000) finds that states that possess an abundant amount of point resources are more likely to experience conflict than states that possess only diffuse resources.

Valuable and easily exploitable natural resources constitute opportunities for rebellions by providing financing for rebel recruitment and arms purchases (Buhaug & Lujala, 2005). Addison et al. (2002) find that natural resources, especially minerals constitute an increase in direct gains for potential rebels and Lujala et al. (2005) find a strong relationship between diamonds and the onset of civil war. Natural resources can play into greed (opportunity) and grievance theoretical frameworks. The natural resources can be used as a source of financing for rebel groups, providing them with opportunity, but can also contribute to grievances because natural resources tend to be associated with ineffective and corrupt regimes (Buhaug, 2005). Resource extraction is mostly spatially fixed, making natural resources extremely amenable to taxing and looting because businesses generally pay the person in power (Buhaug & Gates, 2002).

The empirical evidence for direct connection between natural resource abundance and civil war vary with the operationalization of the resource proxy (Buhaug & Lujala, 2005). Collier & Hoeffler (2004) find strong evidence of a positive association between dependence on primary commodities and the likelihood of experiencing a civil war, but Fearon & Laitin (2003) failed to

reproduce the findings using the same proxy with a different research design. The results demonstrate mixed results, but this could be due to the measurement of natural resources in the literature. Natural resources are typically proxied by the ratio of primary commodity exports to GDP at the state level creating a neglect for the spatial aspect of natural resources. My study will address this weakness by measuring the presence of natural resources at the grid level within states.

Terrain

Most civil war studies include a measure of mountainous terrain and a control for forest cover. Rebel movements prefer to operate from peripheral bases in mountainous or densely forested regions to provide safe havens out of reach from government forces. Dense forest provides cover from aerial detection and hinders the movement of mechanized troops. Mountains offer vital hideouts and also hinder conventional warfare. Many cases demonstrate how military superiority can be counterbalanced by terrain (Buhaug, 2005). Rough terrain is thought to provide an advantage to rebels in civil war, which is why it is modeled in the civil war onset and duration literature.

Physical geography measures are modeled using proxies, causing a gap in how these variables interact with conflict. Rough terrain is usually measured as some approximation of the average share of mountainous and forested terrain or the availability of valuable resource proxies by the ratio of primary commodity exports to GDP (Buhaug & Rod, 2006). Terrain measures have failed to produce robust results, and in many cases produce opposite results from different studies. Fearon & Laitin (2003) conclude that mountainous terrain is significantly related to higher rates of civil war; whereas Collier & Hoeffler (2004) find that in civil war onset, terrain indicators do not produce significant effect.

2.2 Human Geography

The other side of geographical influence on civil wars is human geography, or the concepts of identity and population. Identity, or ethnic and religious groups are believed to be a cause of conflicts. Human geography is considered to be socio-cultural variables, but each concept has a geographic aspect. Where ethnic groups are located, or where the major population centers are located are important to the civil war literature.

Identity

Identity refers to civil wars fought between different ethnic, linguistic or religious groups. A good amount of civil conflict literature focuses on ethnic conflict (Denny & Walter, 2014; Wolff, 2006; Fearon & Laitin, 2004). The theoretical framework of ethnic civil conflict is based on the existing literature on grievances and opportunity. According to Denny & Walter (2014), political power has historically been divided along ethnic lines, making ethnic groups more likely to have grievances against the state, which can make mobilization easier and more probable. Furthermore, Ethnic groups are also more likely to live in concentrated or geographically peripheral areas creating the opportunity to rebel.

The base of support of a rebel group is the foundational link between geography and civil wars. In order to sustain a rebellion and challenge the state, groups must have a base of support. Identity is important to civil conflicts, as it is thought to maintain coherence of the rebel group (Buhaug, 2005); furthermore, a group's identity does not need a long history but rather a common desire (economical, political, or ideological) to create change of the dominant group. Common language and culture facilitate the spread of ideas and information to gauge support (Bates, 1987). Bormann, Cederman, & Vogt (2015) find that intrastate conflict is more likely with linguistic dyads than religious ones. Spatial proximity and dense social networks make it

easier for leaders to identify relevant members and coordinate recruits (Denny & Walter, 2014; Weidmann, 2009; Fearon, 2006; Esteban & Ray, 2008). Ethnic leaders also have an advantage in geography in terms of operating in ethnic enclaves, such as in peripheral and hard-to-reach areas; this theory is closely related to terrain.

Usually measured as distribution of ethnic and religious composition of the population, ethnic composition can be operationalized along two dimensions: fragmentation and polarization or dominance. Fragmentation is the higher probability that two individuals drawn randomly are from different groups (Buhaug & Gates, 2002; Collier & Hoeffler, 2004). Collier & Hoeffler (2004) and Fearon & Laitin (2000) find ethnic fragmentation does not contribute to conflict risk. Dominance occurs if the largest ethnic group constitutes 45-90% of the population. There is broad consensus that dominance is positively related to conflict (Buhaug & Gates, 2002; Collier & Hoeffler, 2004; Ellingsen, 2000; Hefre et al., 2001). This study will build on the ethnic dominance and fractionalization theories and model ethnic groups based on majorities and number of ethnic groups for each district.

Population Size

A state's population and size has extensively been researched in connection to civil conflict. Population size is measured by population density or by the size of the state. In state-level analysis, large states have more civil wars than small states, (Raleigh & Hegre, 2009; Collier & Hoeffler, 2004; Fearon & Laitin, 2003; Hegre & Sambanis, 2006) but there remain several competing reasons for the relationship. Buhaug (2005) explains larger states are considered more prone to internal conflict because larger populations are harder to monitor and control by a central government and because these states contain a higher number of potential

rebel recruits. Furthermore, larger states are more likely to have multiple overlapping conflicts that last a longer time, therefore creating the impression that conflicts last longer in larger states.

Population is positively associated with duration of civil conflicts (Collier & Hoeffler, 2004; Fearon 2004). Studies on distance, such as distance from the capital city is theorized to contribute to the positive association. Sambanis (2004) tested different variables across different coded civil war onset models and found that population size is very robust and significantly associated with civil war. Furthermore, he theorized the threshold of civil war conflicts is 1,000 battle deaths, which is less likely to be reached in smaller states. Raleigh & Hegre (2009) find that conflict events tend to have frequencies in proportion to the size of the population. The findings indicate the locations of the population correlate with the locations of conflict, making population an important aspect for conflict prediction.

2.3 Distance

Building on international conflict studies, some studies have begun to incorporate the role of distance into civil war studies. Boulding's (1962) work on distance decay of power from a state provides the foundation of study of distance in civil war studies; the national strength of a state's home base declines the farther from the state's home base. This concept is applied to intrastate conflict; instead of distance between states or capitals, civil war literature examines distance between a state's capital and a rebel base. Buhaug & Gates (2002) find that identity based, and secessionist wars tend to be located further away from the capital than other types of conflict. Whereas Raleigh & Hegre (2009) explain the effect of distance is moderate compared to population clusters, but states with populations that are largely concentrated around the capital have fewer internal conflict events than countries with populations that are spread out and are also concentrated in locations far from the capital. Salehyan (2009) found that international

borders and safe havens in neighboring states allowed rebels the opportunity to sustain their forces while being less susceptible to government repression. Absolute distance is viewed as a meaningful predictor of government and rebel strengths, public good provision, and political marginalization (Raleigh, Witmer, & O'Loughlin; 2010). Overall, distance is an important indicator in the civil war literature and can be used to locate conflict locations.

Chapter Three: Spatial Analysis and Conflict

This section discusses the role of spatial analysis in the conflict literature. The literature of spatial analysis on conflict consists of interstate, intrastate, and subnational analysis. Through the subfield of spatial econometrics, the quantity of literature linking conflict with space and time has grown. Spatial econometrics provides the foundation to the spatial aspect of this study. Spatial analysis has also allowed for the creation of conflict event data disaggregated with fine spatial resolution. From this data, researchers have been able to test the spatial aspect of conflict.

3.1 Spatial Econometrics

Spatial Analysis of conflict grew out of the framework of 'opportunity' and 'willingness'. Proximity, contiguity, and distance are the theoretical concepts, in which spatial analysis of conflict is rooted. Geographers have explored contiguity trough spatial autocorrelation based on Tobler's first law of geography stating: "everything is related to everything else, but near things are more related than distant things" (Cliff & Ord, 1973; 1981). These three frameworks were used to apply spatial analytical techniques to conflict processes literature. Traditional studies of conflict focus on states at various levels of analysis (global, regional, dyaic), a growing number focus on the conflict locations and the clustering of conflict locations. According to Raleigh, Witmer, & O'Loughlin (2010), the goal of spatially analyzing conflict patterns is to model distinguished influences within qualitative literature, with a quantitative framework. The ability to spatially assess relationships is grounded in the concept of spatial econometrics; a term coined by Jean Paelinck in 1974 (Paelinck & Klaassen, 1979). Anselin (1988) defines spatial econometrics as a subfield of econometrics that deals with spatial interaction (spatial autocorrelation) and spatial structure (spatial heterogeneity) in regression models for cross-sectional and panel data.

Spatial phenomena, such as conflict, exhibits characteristics of both first order and second order spatial effects, whereas the null hypothesis is that the phenomena is spatially random between neighbors. Spatial heterogeneity is the first order spatial effect characteristic of processes that vary systematically over large areas or regions. Spatial dependence is the local-scale effects or clustering of a process; spatial dependence is often measured using spatial autocorrelation statistics (Raleigh, Witmer, & O'Loughlin, 2010; Anselin, 1988; Anselin & O'Loughlin, 1990; 1992; O'Loughlin, 1986). Spatially proximate units are more likely to behave similarly than spatially distant units. The theories predict positive spatial autocorrelation, the spatial clustering of similar behaviors among neighboring observations. Under these theories, spatial dependence may be produced by the diffusion of behavior between neighboring units. As Darmofal (2006) notes, alternatively, neighboring units may share similar behaviors due simply to the units' independent adoptions of the behavior. The spatial dependence observed reflects the geographic clustering of the sources of the behavior in question.

Univariate spatial autocorrelation is diagnosed via global and local measures of spatial autocorrelation. Spatial analysis is measured on two levels of analysis: the global scale effects and local scale effects. The global scale effects are measured utilizing spatial autocorrelation tools such as Moran's I (Moran, 1948) or Getis and Ord's G (1992). The global statistics measure the degree of similarity in the variable of interest between neighboring units to find a correlation between a unit and its neighbors. The local level of analysis employs Local Indicators of Spatial Analysis (LISAs) to identify conflict hot spots, such as Local Moran's I (1948) & Getis Ord Gi* (1992). The global measures of spatial association were expanded upon to create LISAs. Anselin (1995) defines LISA as any statistic that satisfies the following two requirements: The LISA for each observation gives an indication of the extent of significant

spatial clustering of similar values around that observation and the sum of LISAs for all observations is proportional to a global indicator of spatial association. The local measure made way for spatial cluster analysis. The identification of conflict clusters is equivalent to hot spot analysis of crime and disease patterns (Raleigh, Witmer, & O'Loughlin, 2010). Local spatial clusters, or hot spots, may be identified as those locations or set of contiguous locations for which the LISA is significant.

The existing literature on interstate conflict demonstrates that wars are concentrated within space and time. At a global level of analysis, conflict has been demonstrated to cluster regionally. Bremer (1982) finds that the initiation of an international dispute increases the likelihood that another dispute will occur within that region. Kirby & Ward (1987) found significant patterns of spatial association indicated by Moran's I in the relationships between states, their behavior, and war behavior. O'Loughlin & Anselin (1991) used spatial econometric modeling to assess spatial heterogeneity and spatial dependence of national attributes on conflict & cooperation between states in Africa. They found African states interact predominantly with first-order neighbors and that interaction, either conflictual or cooperative, beyond that level is rare and insignificant for the African system as a whole. Both spatial dependence and spatial heterogeneity are strongly present in the African data.

Studies of conflict clustering are rare in quantitative international relations research, but some studies have filled this void. One of the earliest implementations of cluster detection was the geographical analysis machine (GAM) by Openshaw et al. (1988). Houweling & Siccama (1985; 1988) identify spatial distance between and among war outbreaks and utilize a temporal proxy to find battlefield locations of wars clustered in space and time. Braithwaite (2006; 2010a) extends this theory to militarized interstate disputes (MIDs) and finds MIDs cluster in space and

time as well. On the intrastate conflict level, Buhaug & Gleditsch (2008) found that armed conflict tends to cluster spatially in certain geographic areas. Countries in proximity to states experiencing conflict are much more likely to become involved in violent conflict. Geographic clustering of intrastate conflicts strongly suggests that the risk of civil war is not determined merely by attributes of individual countries, and that regional factors and events in neighboring states can alter the prospects for violence. There is a neighborhood effect of civil war where armed conflict in one state makes neighboring countries more prone to violence. Khatiwada (2014) used spatial analysis to aggregate conflict in Nepal at the district level, before using Moran's I to test the spatial dependence of conflict in the districts. Local Moran's I was then run to identify high conflict districts. Conflict clustering literature also includes the level of terrorism events. Nemeth, Mauslein, & Stapley (2014) performed a hot spot analysis to identify local areas of domestic terrorism using a grid-based unit of analysis; they found the variables (mountainous terrain, proximity to a state capital, large population, high population density, and poor economic conditions) increased the likelihood of terrorism. The literature of conflict clustering demonstrates the expectation of the theory extending from wars and MIDs to sub-national political violence events.

3.2 Diffusion

Similar to the concepts of territory, proximity, and distance, diffusion literature grew out of spatial econometrics literature and places the importance of borders as an important variable in determining the frequency of war participation; it is closely related to proximity. The geographic concepts in international relations literature demonstrates that proximate states are similar in their institutional setup (Gleditsch, 2002). According to Schutte & Weidmann (2011), the similarity of spatial proximate outcomes can be explained by two broad categories of

approaches. The first approach assumes spatial clustering occurs because the phenomenon to be explained results from factors which are themselves spatial clustered; the second approach assumes the phenomenon is "contagious", or what is referred to as the "diffusion process".

According to Raleigh, Witmer, & O'Loughlin (2010), early work in the spatial analysis of conflict focused on how neighboring states influence the tendency for international disputes to spread, expanding on Most & Starr (1980) concept of "opportunity and willingness". Most & Starr (1980) presented two spatial hypotheses of diffusion. (1) Positive spatial diffusion, or the process in which the occurrence of a new war participation in a state increases the likelihood that other states will experience subsequent war participation, and (2) Negative spatial diffusion, if occurrences of war participation decrease the likelihood of subsequent war participation by others. Starr & Most (1983; 1985) later test their hypothesis on spatial diffusion on Africa and found the results in Africa are similar to those on the global level. They also found the type of war did not affect the diffusion across borders. According to Diehl (1991) much of the other scholarly diffusion literature relied on the analogy between the spread of disease and the spread of war. The theory is based on the literature of disease diffusion. When a person gets infected, the disease tends to spread to people near the source of the infection.

Related to diffusion and borders, other diffusion literature focuses on the extent war can spread. Bremer (1982) first suggested geographic distance was a limiting condition for how far wars can spread. Faber, Houweling & Siccama (1984) explain that outbreaks of war in specific regions are not related to outbreaks in other parts of the international system. Houweling & Siccama (1985) then used an epidemiological approach using the spatial and temporal components of interaction to find again that diffusion is confined to certain geographic areas. This approach is significant because the incorporation of distance provides the foundation for

neighborhoods in spatial analysis. Siverson & Starr (1991) confirm the probability of war diffusion is substantially increased as opportunities and willingness increase while using borders and alliances as indicators.

Literature on intrastate conflict and diffusion have been inconclusive. According to Raleigh, Witmer, & O'Loughlin (2010), this is partly due to inconsistent empirical specifications, which include different data sets, varying definitions of conflict and explanatory variables, use of different spatial weighting schemes and time periods. Schutte & Weidmann (2011) attempt to fill in this void by exploring the spatial-temporal dynamics of violence within a single conflict, rather than its spread between state borders. They differentiate between two patterns of diffusion: relocation and escalation. Nevertheless, levels of democracy in neighboring states tends to have an indirect influence on civil war risk. Sambanis (2001) found that lack of democracy in a neighboring state with an ethnically divided society can have a negative effect on the domestic ethnic antagonism. Gleditsch (2002) also found that democratizing states located near democratic states have a lower risk of experiencing a civil war. This is further supported by Gleditsch (2007), which found transnational linkages between states and regional factors strongly influence the risk of civil conflict, suggesting risk of civil war is related to a state's linkages. Still, it remains unclear whether high neighboring civil war risk is related to unstable and poorly controlled border regions, rivalry across neighboring states, or to similar socio-economic conditions (Raleigh, Witmer, & O'Loughlin, 2010; Murdoch & Sandler, 2002; Salehyan & Gleditsch, 2006).

3.3 Spatial Analysis of Conflict

Geographic Information Systems (GIS) has allowed researchers the ability to analyze spatial relationships in international relations more advanced than traditional studies. The

advances of GIS have placed studies into two broad categories: studies that focus on the locations of conflicts and studies that focus on the violence within conflicts. Studies on the locations of conflicts include studies that create geospatial datasets of conflict locations and studies that analyze the locations of conflict locations.

There are two types of studies that deal with the locations of conflict. The first one is studies that produce geospatial data sets involving conflict. According to Branch (2016), GIS has made it possible for scholars to develop new data sets and variables, disaggregate data to more theoretically or empirically appropriate levels, merge data in new ways, and test novel spatial hypotheses. Event datasets grew out of the necessity of geospatial datasets with the advances of geospatial analysis. Traditionally, the Correlates of War project served as the main supplier of data used for spatial studies. Militarized Interstate Dispute Location (MIDLOC) dataset is housed in the Correlates of War project; it provides details of the geographic location of Militarized Interstate Dispute onsets between 1816 and 2010 (Braithwaite, 2010b). The MIDLOC dataset deals with interstate conflict data; there are a number of civil conflict datasets. The UCDP/PRIO Conflict dataset provides a dataset on conflicts that do not reach the level of a militarized interstate dispute, lowering the threshold to 25 battle deaths (Gleditsch et al., 2002). UCDP also provides a disaggregated spatial / temporal data of organized violence (Sundberg et al., 2013). Africa receives a great deal of attention from quantitative spatial datasets. There are two that focus on Africa: Social Conflict Analysis Dataset (SCAD) and Armed Conflict and Location Event Data (ACLED). SCAD focuses on conflict types not usually focused on, such as protest, riots, and other forms of social conflict not systematically tracked (Salehyan et al., 2012). ACLED marks the actions of rebels, governments, and militias within unstable states, specifying the exact location and date of battle events, allowing research on local level factors

(Raleigh et al., 2010). The ACLED dataset is used in this study because it is the most suitable dataset for sub-national conflict analysis. The creation of spatially-oriented conflict datasets has allowed for the studies on conflict locations and spatial analysis of conflict.

The other type of research focuses on the locations of conflicts within states. According to De Juan (2012), these studies are further subdivided into focus on conflict onset or incidence and those interested in intensity or duration. The literature that examines conflict onset or incidence explores the idea of battlespace or 'ConflictSpace'. Flint et al. (2009) produced a systematic analysis of interstate conflict data by modeling the spatiality of conflict through spatial analysis and the territorial and network embeddedness through social network analysis. ConflictSpace is defined as incorporating the analysis of multiple spatialities into the analysis of war (Leitner, Sheppard, & Sziarto, 2008). The results demonstrated the spread of World War I through the concept of ConflictSpace and locations of interstate conflict. On a subnational scale of analysis, Buhaug & Rod (2006) used GIS and logit modeling to identify regions of conflict and to generate subnational measures of key explanatory variables. Sparsely populated regions near the state border, away from the capital, and without significant rough terrain make territorial conflicts more likely. Whereas, regions that are densely populated, near diamond fields, and near the capital city, make governmental conflicts more likely. Their results reveal spatial clustering of conflict that covaries with the spatial distribution of several related factors. Furthermore, Theisen et al. (2012) examined how precipitation data is related to conflict onset. They find no support to the theory conflict is likely to break out in areas affected by drought and water scarcity. Derived from the literature on strategic locations, Hammond (2018) uses GIS and network analysis to operationalize strategic locations based on population settlements and road

networks. He finds that during conflicts, locations with control access to other areas within the state are significantly more likely to be fought over.

The studies examining the intensity or duration of the locations of conflict are fairly limited. Buhaug et al. (2009) finds that conflicts located at considerable distance from the main government stronghold, along remote international borders and in valuable minerals area last longer. Lujala (2009) explores the overlap of conflict zones and the abundance of natural resources on the intensity of conflicts; finding that conflicts fought in areas with a plethora of natural resources are more severe than others. Furthermore, Rustad et al. (2011) examines the role of duration in the geographic overlap between conflict areas and forested terrain. They do not find conflict duration is increased due to the overlap.

According to De Juan (2012), most subnational geospatial analysis focuses on the location of violent events within conflicts. The literature can again be subdivided into conflict onset, incidence & occurrence and that which discussing intensity or duration. Studies about the occurrence of conflict events explore which subnational units have higher or lower risk of experiencing violence. Hegre & Raleigh (2009) explore different geographic factors that influence the probability that subnational areas will experience violent events. They find that population concentration and distance to capitals and borders impact the risk of event occurrence. Theisen (2012) applied a similar approach while analyzing areas of low levels of land per capita, which displayed a higher risk of experiencing violence within Kenya. He does not find evidence to support his thesis. In the North Caucasus region of Russia, O'Loughlin & Witmer (2009), find violent events are more likely to occur in regions close to the highway, with a low share of ethnic Russians, and with forest cover.

The other subdivision deals with how conflict intensity and duration are distributed geographically. In the civil war in Nepal, Murshed & Gates (2005) and Do & Iyer (2010) examine the geographic variation of conflict intensity measured by the number of conflict-related deaths per district. Both studies found poverty and inequality contribute to conflict-related deaths. Costalli & Moro (2011) explore how ethnic settlement patterns influenced the severity of fighting during the Bosnia Civil War at the municipality level; finding that ethnic fractionalization and polarization are important. The literature of conflict intensity demonstrates an importance on modeling subnational poverty and ethnic fractionalization.

3.4 Issues with Spatial Analysis

The need of spatial analytical techniques in international relations literature grew out of the weakness in traditional studies to conceptualize the spatial aspect. According to Braithwaite (2010a), analysis of conflict locations allows to more accurately assess the spatial distribution of conflict. Furthermore, a major reason for integrating GIS and the study of conflict is to facilitate data generation on a truly sub-national level, as opposed to relying on national statistics. But, despite the potential gains of the fairly untapped toolset of spatial analysis of conflict, caution must be taken. There are two primary challenges associated with the application of GIS to international relations: Measurement validity and selection bias.

Scale of measurement of geographic indicators is one of the biggest concerns with the validity of spatial analysis. Measurement validity is undermined when institutions, practices, ideas, or behaviors are not accurately described by the points, polygons, or pixels used in the GIS data set (Branch, 2016). Associated with measurement, Buhaug (2005) states, the theories and hypothesis about geography and conflict often speak of sub-national conditions but the variables are usually tested by estimating the statistical association between a states resource wealth and

the involvement in a civil war. Measurements of variables such as rough terrain and natural resources might be present in a state experiencing a civil conflict, but these variables are not spatially associated with the conflict incidents, which could give a false sense of causal relationships.

The issue of the modifiable areal unit problem is related to the issue of measurement validity. The modifiable areal unit problem refers to the fact that results of spatial statistical analysis are sensitive to a zoning system used in data aggregation (Fortheringham & Wong, 1991; Rogerson, 2015). Inferential models are hindered by the lack of information for predictors at the level of disaggregation that match conflict data. In other words, if subnational patterns of violence do not match the subnational patterns of the independent variables, it is difficult to draw reliable conclusions on the actual effects of the factors under investigation (Raleigh, Witmer, & O'Loughlin, 2010). The two most common methods of aggregating conflict data are either local level administrative districts or a grid-based system; the size and location of the grids can alter the results of the statistical analysis. Therefore, it is imperative to examine the sensitivity of the results to modifiable areal units.

Chapter Four: Conflict and Machine Learning

This section discusses conflict prediction methods and limitations, rather than focusing on the theoretical framework of the models and methods. First, a lack of consensus remains in the prediction and forecasting (used interchangeably) literature on the importance of theory in conflict predictions. According to Jones & Linder (2015), machine learning methods deliver good predictions, yet are not very useful for theory driven work. Second, the theoretical foundation of this project is grounded in the fields of geography and quantitative international relations. Prediction models have grown remarkably since the 1960s when prediction models were based on simple linear-regression models. According to Witmer et al. (2017), predictions are based on two general motivations. One uses predictions to assess the influences of independent variables upon the outcome of interest, the other uses modeling and simulation techniques to forecast observed trends into the future. Conflict prediction utilizes standard machine learning methods, such as time-series, logit, and neural networks; but the literature that combines spatial analysis and conflict prediction remains very small.

4.1 History of Conflict Prediction

Despite the advances in machine learning in the era of "big data", conflict prediction is not a new concept—it dates back to the 1960s. Hegre et al. (2017) breakdown the history of prediction in peace research literature into three generations of studies. The first generation of conflict prediction was inspired by Richardson (1960), who created a systematic collection of data on wars to predict their occurrence. His book is considered the founding book of the quantitative study of conflict (Ward et al., 2013). Inspired by the earlier work of Richardson, Wright (1942), and Sorokin (1957), the Correlates of War project was founded with the goal of systematically accumulating scientific knowledge about war. According to Ward et al. (2013), it

provided the gold standard of quantitative research in the twentieth century. The first generation of conflict prediction is characterized by the efforts to collect scientific data on conflict to produce early-warning systems. Early events-data projects also highlighted the capabilities of forecasting and provided a template for collecting fine-grained data effective to approximate real-time conflict early warning. Conflict prediction was focused on interstate conflict, as interstate conflict was the focus of international relations literature.

The second generation of conflict prediction came in the early 1980s after nearly a decade of little attention. According to Hegre et al. (2017), the second generation of conflict prediction produced two critical innovations. The first innovation was the work of Bueno de Mesquita (1980), which used the expected utility theory, mentioned in section one, to provide a link between theory and conflict prediction. Another innovation was the ability to build statistical models based on news source data; Philip Schrodt (1988) pioneered the use of automated data and event coding. The surge in event data in the 1990s is a result of the increase in computing power that made it feasible. Conflict prediction was focused on interstate conflict but Schrodt's innovation provided the foundation for researchers to address lower level conflict that did not reach the threshold of the Correlates of War Project. The data used for this project derives its foundation to Schrodt's work on automated data and event coding. Schrodt also changed the format of conflict analysis from country-year to conflict-year. The machine-coded data was first introduced by Schrodt, Davis & Weddle (1994), when they used algorithms to automatically classify and code political events based on news articles. The Kansas Event Dataset (Schrodt, Davis, & Weddle, 1994) and Integrated Conflict Early Warning System (ICEWS; O'Brien, 2010) were among the first machine-coded datasets.

As event data became more fine-grained, the policy community became interested in early warning systems. The third generation of conflict prediction saw it grow into a subdiscipline of conflict research. According to Hegre et al. (2017), conflict prediction is now seen as a "mainstream" effort by the wider scientific community. The United States government began to sponsor different programs, such as the Political Instability Task Force, with the goal of predicting political instability two years before it occurs. According to Ward et al. (2013), growing demand led to the development of research and dataset on the micro-level and the advances of GIS allowed the data to be disaggregated spatially and temporally. The three generations of conflict prediction provide the foundation for the different methods discussed below.

4.2 Conflict Prediction Methods

The first prediction models relied on simple linear-regression models then evolved to machine learning techniques, such as neural networks and random forest decision trees. Random forest decision tree methods are becoming more popular with the continued advances in machine learning and artificial intelligence because they are simple and useful for interpretation. Random forest is a type of tree-based method, along with bagging and boosting. First proposed by Breiman (2001), the Random Forest algorithm consists of many smaller models. Predictions are obtained by combining the outputs at all the smaller models; the smaller models are classification and regression trees (Jones & Linder, 2015). Classification and regression trees rely on repeated partitioning estimating the conditional distribution of a response given a set of explanatory variables.

This study is rooted in the body of literature on random forest decision tree methods. Hill & Jones (2014) were an early use of random forest in international relations literature; they used

random forest to determine the predictive power of measures of causes of state repression. Random forest has since been used in comparison with other machine learning models for civil war forecasting. Blair & Sambanis (2016) use a random forest model to demonstrate how a "process-based" model outperforms a-theoretical alternatives and outperforms models based on a state's structural characteristics. They find their theoretically-driven model generates accurate forecasts. Muchlinki et al. (2015) examined civil war onset data comparing a random forest with a logistic regression. Logistic regression is a poor predictor of civil war onset. It is likely that the predictions of the logistic regressions will be biased toward the majority class because civil war onset data tends to be unbalanced. Two approaches have been developed to correct the bias: rare event logistic regression and L1-regularized logistic regression. They found random forest can learn complex patterns in the class-imbalanced data and more accurately separate instances of peace from instances of conflict onset better than logistic regression models.

Conflict early warning systems and subnational analysis also employ random forest techniques for conflict prediction. Perry (2013) uses machine-learning methodology to model fragility and vulnerability to conflict at the global level; he compares two machine learning models: naïve Bayes and random forest. He finds the random forest algorithm offered better results than the naïve Bayes algorithm; one advantage of the random forest is the ability to determine variable importance based on each variable's contribution to an increase in purity and a decrease in error. The ability to rank the importance of the variables is an important step to assess the importance of the variables being modeled spatially in this study. The spatial model of this study should identify how variables interact with conflict locations based on a spatiality of the variables. Lazicky (2017) predicts conflict at the subnational level, as opposed to the cross-national level, and is one of the few studies that applies spatial analysis to predict conflict. He

assesses three machine learning models: Logistic regression model, Lasso model, and random forest and a spatial model, geographic weighted regression and finds the conflict forecasting model should incorporate a hybrid of the lasso and logistic models. Yet, he only assessed case studies of UN peacekeeping missions and examined only violence against civilians. Therefore, his model only tested a specific issue and not a broad conflict in the region. Another study at the local level, Blair, Blattman & Hartman (2017), applies machine learning techniques, lasso, random forest, and neural networks, to panel survey data in Liberia to predict violence two years in the future. Their findings demonstrate lasso technique is the best at predicting local violence based on survey data. While Lazicky (2017) and Blair, Blattman & Hartman (2017), find that the random forest model is not the best model for conflict prediction, the random forest model is better suited for the spatial analysis approach of this study.

The most common method of conflict prediction is the use of a logistic (logit) regression model. Conflict data is typically coded using a binary method (0,1) with the event either taking place within a conflict-year or not; the logistic regression is most suitable for a binary dependent variable. Ward & Gleditsch (2002) emphasizes spatial information as predictors of conflict and compares a logit model with Markov chain Monte Carlo (MCMC) technique. They find the unconditional logit model and pseudolikelihood estimates of the autologistic model do not yield predicted probabilities above 0.5 for any observation, but the predicted probabilities from the MCMC estimates of the autologistic model exceed the 0.5 threshold. Logit models have been used at different level of analyses, and they have been used successfully at the subnational level. Weidmann & Ward (2010) created a logit model to make predictions at the municipality level for the Bosnia Civil War. They forecasted four municipalities correctly but missed three actual outbreaks and over-predicted four other municipalities. Despite the accuracy, it was one of the

first studies that mapped the results of the prediction model. Weezel (2017) used a logit model for subnational data in Africa from 2000 to 2009; out of sample predictions from 2010 to 2015. He found the strongest predictor of future conflict is current conflict incidence in the grid-cell and neighboring cells. This study further tests that theory by creating different types of conflict clusters and its effect on the random-forest based decision tree model. Using the logit model, Weezel (2017) furthermore found that most geographic factors have low predictive power, but travel time to the nearest city is a strong predictor of future conflict. Also, no strong conclusions can be drawn whether this is because of a structural difference between urban and rural areas that drives conflict incidence or whether this result is driven by potential reporting bias in the conflict data. Celiku & Kraay (2017) compare the performance of a binary regression classification algorithm and random forest versus two unconventional classification algorithms: linear & threshold classifiers, exploring conflict and non-conflict episodes using a country-year approach. They found the threshold classifier had the best overall predictive performance.

Neural networks were among the earliest methods of incorporating machine learning and artificial intelligence in conflict research. Neural networks gain attention because their structure resembles the structure of biological nervous systems. One of the earliest examples of machine learning for conflict prediction using a neural network was Schrodt (1991b), who compared the accuracy of a neural network model to a discriminant analysis, logit analysis, and a rule-based ID3 algorithm. He found the neural network outperformed both discriminant analysis and rule-based ID3 in terms of accuracy but was roughly similar in accuracy to multinomial logit analysis. Beck, King, & Zeng (2000) compare the predictive power of logit model of conflict with a neural network and finds neural networks have a better ability of capturing the nonlinearities and interactions driving the incidence of conflict. Lagazio & Russett (2003) argument that neural

networks provide an alternative to multivariate statistical techniques confirms a Beck, King, & Zeng (2000) analysis. But neural networks tend to not consider geography and are not suitable for a spatial conflict prediction.

Time series methods are another technique for conflict prediction. The most common technique for employing time series methods is vector auto-regression (VAR) models. Time series models were first used to model the sequence of cooperative and conflictual events (Goldstein, 1992; Schrodt & Gerner, 2000). Brandt & Freeman (2005) and Brandt, Colaresi & Freeman (2008) use event data from the Kansas Events Data System to demonstrate how Bayesian time series models, such as Bayesian vector auto-regression (BVAR) and Bayesian structural vector auto-regression (BSVAR), can be used to analyze conflict dynamics and make ex post forecast in the Middle East over the short term. Brandt, Freeman, & Schrodt (2014) builds off the previous work to provide near real time forecast for conflict between Israelis and Palestinians for 2010. Yonamine (2013) builds off this work to implement an Autoregressive Fractionally Integrated Moving Average (ARFIMA) model, which models all univariate time-series independently of each other. Using GDELT data to model a time-series forecast that explores the district month level within Afghanistan, he fines the model performs better than the baseline, and empirical forecast of violence should use as finely-grained geospatial aggregations as possible. Yonamine (2013) time-series approach was the first time-series method that utilized cartography, or maps, as a method of viewing the results.

Hidden Markov Models (HMM) are a less common method of conflict prediction and closely related to time series models. HMMs are a sequence comparison method as a computationally efficient method of generalizing a set of sequences observed in a noisy environment. In other words, they are a quantitative pattern recognition technique that compares

an existing sequence of behaviors to a set of similar historical cases. They have been demonstrated to capture escalations of conflict and use the patterns to forecast future escalations (Schrodt, 1999; Schrodt, 2000; Shearer, 2007; Schrodt, 2006; Schrodt et al., 2013; Bond et al., 2004). The models have been used in the Middle East to forecast escalation and outbreak of conflict. Schrodt (1999; 2000) used HMMs to forecast the outbreak of armed violence between Israel and Arab forces in south Lebanon from 1979 to 1997 (excluding 1982-1985). They use six training cases of "tit-for-tat" escalation before fitting for the time frame. They find the model identifies about half the conflicts, making it suitable as an event-based monitoring system but ineffective as an early warning indicator. Furthermore, he finds HMMs could be used to develop conflict measures based on the event similarities to historical conflicts rather than on aggregated event scores. These findings provide the rationale to incorporate emerging hot spots as a variable in my random forest model. The emerging hot spot will identify past conflict locations, and those locations can be used to inform the prediction model. While HMMs have been an effective pattern recognition technique, the model lacks the ability to model events spatially, thus making it not a suitable method for a spatial analysis approach.

Other less common, but noteworthy alternative approaches include geographic techniques such as point processes, cluster analysis, and climate-sensitive models. Schutte (2016) uses a point process model to predict major conflict zones across civil conflicts in 10 case study states within sub-Saharan Africa. He finds the locations of high-intensity conflict zones are correctly predicted six out of 10 times. The point process model was chosen to model intensity, as opposed to occurrence. Schutte (2016) also claims high intensity conflict regions yield more than 50 percent of the maximum conflict intensity compared visually to empirically observed hot spots. Witmer et al. (2017) forecast conflict using multiple future climate scenarios and sociopolitical

factors, such as population size and political rights, joined with temperature anomalies. They present multiple forecasts of violence under alternative climate change scenarios, such as optimistic and current global trajectories, political rights scenarios, improvement and decline, and population projections between low and high fertility. The models display that a growing population and rising temperatures will lead to higher levels of violence in sub-Saharan Africa if political rights do not improve. Witmer et al. (2017) predictions are closely related to the long-term forecast funded, the Political Instability Task Force. Basuchoudhary et al. (2018) attempts to build a unified theory to conflict prediction and finds that the game theoretic rational choice models of bargaining and commitment failure predicts conflict better than alternative approaches.

4.3 Potential Issues with Conflict Prediction

Similar to any scientific method, conflict prediction is not without its limitations. Conflict prediction has a few limitations. Schrodt (2014) describes the seven biggest errors in quantitative political analysis as: models that ignore the effects of collinearity; pre-scientific explanation in the absence of prediction; excessive reanalysis of a small number of datasets; using complex methods without understanding the underlying assumptions; misinterpreting frequentist statistics as Bayesian; failure to consider alternative structures to linear statistics; confusing statistical controls and experimental controls.

Conflict events are considered rare events, which makes them difficult to sample and create an unbiased model. As stated earlier, Muchlinki et al. (2015) find civil war onset data tends to be unbalanced; the ratio of conflict years to peace years is roughly 1:100. King & Zeng (2001) further explain rare events are difficult to predict because popular statistical methods, such as logistic regression, can shapely underestimate the probability of rare events and

commonly used data collection strategies are inefficient for rare data events. Their solution is to alter the sampling design, such as sampling all events and only a tiny fraction of nonevents.

Conflicts are often dynamic, and the complexity of the international system is difficult to model. According to Cederman & Weidmann (2017), the most pernicious problem pertains to the failure of studies to appreciate the fundamental complexity surrounding the conflict process. The conflict processes usually involve actors engaging in surprising or rule-breaking ways to accomplish their goals, which makes the prediction environment unstructured. Chadefaux (2017) expands on this issue, state actors do not always respond to different situations the same way; different sequence of events can lead to different outcomes. Modeling the location of conflict events may help mitigate the issue of dealing with various actors.

Another issue with conflict prediction remains on how to assess the results of the models. Ward, Greenhill, & Bakke (2010) find that studies pay too much attention to finding statistically significant relationships, and too little attention to finding variables that improve the ability to predict civil conflicts. Furthermore, Brandt, Freeman, & Schrodt (2014) share the sentiment that methods such as root mean square error and other point metrics are not adequate measures of forecast models. As Celiku & Kraay (2017) explain, conventional approaches for conflict prediction suffer from an internal inconsistency: the objective function that is optimized when the prediction model is fitted to the data is different from the objective function that typically is used to evaluate the quality of the resulting prediction. In random forest methods, the branches in the underlying classification trees are chosen to maximize the homogeneity of observations in the resulting subgroups, weighing all observations equally. Thus, they find support that out-of-sample predictions improves random forest classifiers. Ward, Greenhill, & Bakke (2010) also find that a focus on out-of-sample prediction helps guard against the inclusion of long lists of

explanatory factors that may worsen predictive performance. This study incorporates out-of-sample predictions to assess the accuracy but also maps the predicted values versus actual values. Cartography is a very effective method of displaying results and will be easy for policymakers to understand the results without having to understand algorithms or confusing graphs.

Chapter Five: Methodology

The role of geography is understudied in quantitative international relations literature, especially in the subfield of conflict prediction and forecast. The combination of spatial analysis and machine learning in a single model can yield powerful results for prediction. In this study, a spatial Classification Random Forest model is tested against a spatial Logistic Generalized Linear Regression model to test the predictive power of future conflict events. Spatial analysis is implemented into the models through geoprocessing the explanatory variables and the utilization of distance features in the models. The research design takes influence from the subnational conflict prediction literature (Perry, 2013; Schutte, 2016; Weezel, 2017). Perry (2013) proves the feasibility and initial methodology of using random forest models for conflict prediction at the administrative two level across Africa. Schutte (2016) uses a point process model instead of traditional machine learning methods to predict conflict zones based on geographic conditions. Weezel (2017) uses a logit model to examine geographic and socio-economic factors at the sub-national level. This study will build off the previous methodology to compare the two prediction models using binary conflict occurrence within administrative two districts between 2015 – 2017 as the dependent variable. Twenty-three explanatory variables are used to predict the occurrence of future conflict within the districts, divided into three types: explanatory variables, distance variables, and raster variables . Out of sample conflict occurrence from 2018 is used to assess the predictive performance of the two models; the *F*-1 scores are compared, and the variables are reviewed to identify important variables and gaps for future research.

5.1 Theoretical Background

Spatial analysis is utilized on the explanatory variables to provide the values over space, as opposed to relying on ratios of geographic features. The spatial analysis of this study will be

done using the software package of ArcGIS produced by Environmental Systems Research Institute (ESRI). Spatial features are coded either as vector data or raster data; this study utilizes both data types as variables. Vector data is visualized by points, lines, and polygons. The latitude and longitude of the conflict event data create a point feature at a single location. A polygon is a feature that covers a definable spatial area. Raster data uses pixels, or predefined equivalent-sized units, that assign a value for a signal variable across the size of the data; an example is elevation data or land cover types. ArcGIS allows for both data types to be analyzed simultaneously. The Random Forest model can ingest rasters and distance features directly into the model, which provides an advantage over alternative methods. The Generalized Linear Regression model can use distance variables as an input. Some of the explanatory variables are created using spatial analysis, such as road density and conflict density.

The intent of this paper is to combine spatial analysis with machine learning methodology. Supervised machine-learning is used to produce the conflict prediction. Supervised machine-learning algorithms make predictions from existing data; it uses the existing sample data that is known to the dependent variable to extrapolate onto data that the dependent variable is unknown (Perry, 2013). The Random Forest model was selected as the type of supervised machine-learning algorithm for this study, in part because of its ability to handle a large number of observations and variables and its past performance in the previous literature.

The Random Forest model used is based on Leo Breiman's (2001) Random Forest algorithm. The model is chosen over other models, to test against a logit model because random forest performed better than a logit model (Perry, 2013) and Weezel (2017) claims more sophisticated statistical techniques than a regression model could improve predictions. Random Forests are a form of ensemble learner that extends the decision tree learner algorithm. Ensemble

methods refer to the use of multiple models that are then recombined to increase the predictive performance over any of the single models (Perry, 2013). Decision trees create a model to predict the value of a target variable based on a range of input variables. The trees are composed of a root, children and leaf elements. Each leaf element corresponds to the value of a target variable given all the values of the elements as you traverse the path from the root to the leaf. Each split corresponds to a given input variable and each child element corresponds to that variable's possible values. The tree is "learned" by deriving splits based on some test that determines which input variable best splits the data at that level. Random forests expand on decision trees by iterating over the data to create multiple trees. Each tree is derived from a sub-sample of the training data and a sub-selection of the explanatory variables. The final classification is determined by an aggregation of "votes" from each of the component trees (Breiman, 2001; Perry, 2013).

According to Breiman (2001) Random Forest are defined as "a classifier consisting of a collection of tree structured classifiers: {h(x, Θk), k = 1, . . .}, where the {Θk} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x." The algorithm is given as:

1. For B_i, $i = 1, \ldots, B$
    a. Draw a bootstrap sample S of size N from the training data.
    b. Grow an unpruned random-forest tree, T_b using the bootstrapped data, until the minimum terminal node size, n_{min}, is obtained by recursively following the sub-algorithm.

    Randomly select p variables from total set of P variables.
    Select the optimal variable/split-point among the p variables
    Split node into two daughter nodes.
2. Output ensemble of trees $\{T_b\}_1^B$
3. Predict new observations, or out-of-bag observations

For classification: Let $\hat{C}_b$(x) be the class prediction of the b^{th} random forest tree
→ $\hat{C}_{rf}^{B}(x)$ = majority vote $\{\hat{C}_b(x)\}_1^B$

(Breiman, 2001; Siroky, 2009)

In order to prevent overfitting, the algorithm reduces dependence between trees by making majority voting. M_{try}, the number of variables to try at each node, is the main tuning parameter, using M_{try} = p/3 for classification (Breiman, 2001). Randomly selected explanatory variables can reduce competition between similarly important factors or factors that fit some observations poorly and some well. Random Forest use out-of-bag samples as a built-in cross-validation method (Breiman, 2001). This helps assess the accuracy of the model. Other tuning options include specifying the number of trees and the number of randomly sampled variables used for any given tree in the forest. For classification models, the number is chosen by the square-root of the total number of variables (ESRI, 2019).

The Random Forest algorithm provided by Environmental Systems Research Institute (ESRI) is unique because while the Random Forest model is not inherently a spatial model, the incorporation of distance measures and the ability to read raster layers allows for the incorporation of spatial analysis into the prediction model. The algorithm separates the explanatory variables into three categories: Explanatory Training Variables, Explanatory Training Distance Features, and Explanatory Training Rasters. The distance features and raster data types allow for geographic features to be easily digested into the model.

The random forest model is compared against a logistic generalized linear regression. A logistic model is used to predict the presence or absence of a conflict event in the districts. The Generalized Linear Regression model is based on Nelder & Wedderburn (1972) algorithm. The Generalized Linear Regression has three model components: an exponential family of probability distributions for *Y*, the linear predictor of $X\beta$, and the link function that relates the linear predictor to the mean of the distribution. The algorithm is given as:

The mean, μ, of the distribution depends on independent variables, X, through:

4. $E(Y) = \mu = g^{-1}(X\beta)$

where E(Y) is the expected value of Y; $X\beta$ is the linear predictor, a linear combination of unknown parameters β; g is the link function.

The probability of distribution for *Y* utilizes a Bernoulli distribution to account for the binary data. Given as:

If X is a random variable of the distribution, then:

5. $\Pr(X - 1) = p = 1 - \Pr(X = 0) = 1 - q$

The probability mass function f of the distribution, over possible outcomes k, is

6. $F(k;p) = p^k(1 - p)^{1-k}$ for $k \in \{0,1\}$

The linear predictor:

7. $\eta = X\beta$

Link Function:

8. $X\beta = \ln\left(\frac{\mu}{1-\mu}\right)$

(Nelder & Wedderburn, 1972; Menard, 2002; ESRI, 2019)

Similar to the Random Forest model, the Generalized Linear Regression algorithm is provided by ESRI. Unlike the Random Forest algorithm, the Generalized Linear Regression algorithm does not incorporate raster data types into the model, although it does process distance features. The algorithm splits the explanatory variables into two categories: Explanatory Training Variables and Explanatory Distance Features. The Explanatory Training Rasters from the Random Forest Model are converted into Explanatory Training Variables.

5.2 Performance Metrics

There are a number of different methods to assess the performance of the models. *F*-1 score is the primary indicator of performance for this study. The *F*-1 score is the weighted average of Precision and Recall (Sensitivity). The *F*-1 score is a product of a Confusion Matrix, which takes into account True Positive, True Negative, False Positive, and False Negative values. Figure 1 displays a Confusion Matrix.

Figure 1: Example Confusion Matrix

	Predicted Class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

(Joshi, 2016)

A True Positive is a value that is a correctly predicted positive value and the actual value in the out of sample data is also positive, or districts where the model predicted a conflict occurrence and a conflict occurrence was present in the 2018 out of sample conflict data. A True Negative value is the same except areas where no conflict occurrence was predicted or happened in the district. False Positives are areas where the actual out of sample data displayed no conflict occurrence, but the model predicted an event and False Negatives are areas where conflict occurrence was present but not predicted (Joshi, 2016).

Since the *F*-1 score is the weighted average of Precision and Recall, both metrics are calculated using the four values above. Precision is the ratio of correctly predicted positive values to the total predicted positives values, or Precision = True Positive / True Positive + False Positive. Sensitivity is the ratio of correctly predicted positive values to all the values in the positive occurrence class, or Sensitivity = True Positive / True Positive + False Negative. The *F*-1 score is then calculated using the following equation:

$$F\text{-}1 = 2*(\text{Sensitivity} * \text{Precision}) / (\text{Sensitivity} + \text{Precision})$$

An *F*-1 score is calculated for both models versus the out-of-sample to provide a better accuracy score of the two models.

One of the advantages of using the random forest model is the calculation of the top variable importance (Perry, 2013). Importance is calculated using Gini coefficients, which can be

thought of as the number of times a variable is responsible for a split and the impact of that split divided by the number of trees. Splits are each individual decision within a decision tree (ESRI**,** 2019; Breiman, 2001). The MCC score is measured -1 to 1, with 0 being the model performance equal to random.

The Logistic Generalized Linear Regression model will produce two outputs to interpret the results of the model. The first output used to interpret the results is percent deviance explained. This is the proportion of the dependent variable variance accounted for by the explanatory variables (ESRI, 2019). The second output is a Joint Wald Statistic, which is a measure of overall model statistical significance.

5.3 Dependent Variable

With the help of GIS and faster computing power, conflict research data has come a long way to provide robust sub-national conflict locations. This study uses Armed Conflict and Location Event Data (ACLED; Raleigh et al., 2010) to provide the dependent variable for this study. The ACLED project collects event data that is coded by researchers primarily from secondary source information from news reports and attributed by dates, actors, types of violence, and locations of all reported political violence and protest events. ACLED is used over alternative sub-national conflict datasets because ACLED provides conflict events that occur within civil wars and periods of instability; the data is more robust than just civil war events or terrorism events. This study will examine conflict more broadly, in the form of political violence, than only examining civil war onset. Political violence prediction is modeled because political violence will capture all threats to the population, not just violent events related to wars. Political violence includes battles from civil wars and violence against civilians, such as terrorism attacks. Perry (2013) states that going forward, models should use an expanded definition of violence to

create a better dependent variable. Protest events, headquarters established, and strategic developments events are excluded from the analysis to model direct political violence within the study area.

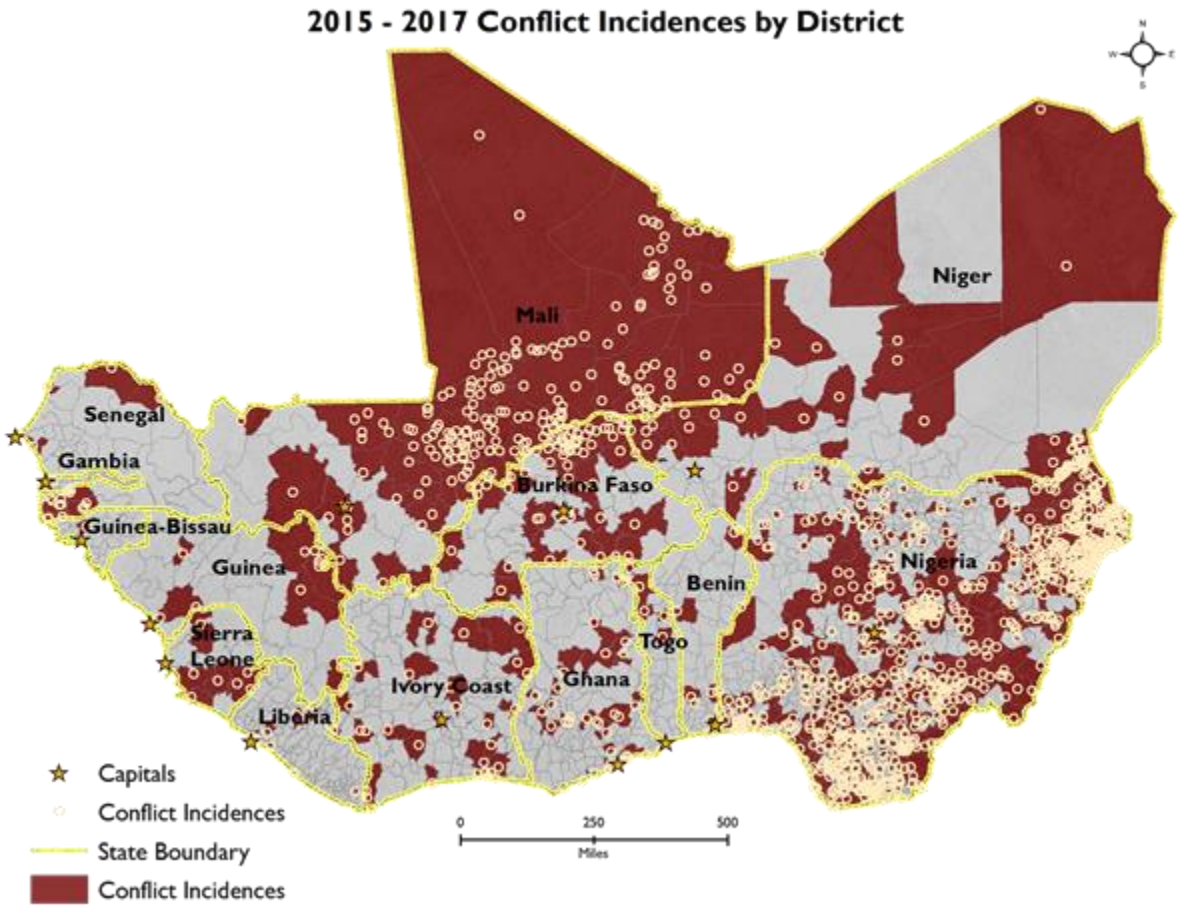
The timeframe of the conflict incidents is 2012 – 2018. Conflict event data from 2015 – 2017 is used to train and perform in sample tests of the spatial random forest model. Events from 2012 – 2014 are used as previous event locations. The time frame of 2018 will provide the out of sample data for the Random Forest classification model and Logistic Generalized Linear Regression model, where the results will be analyzed using traditional machine learning performance metrics, such as Variable Importance and *F*-1 Score. The actual conflict events from 2018 will be mapped against the predicted events from the model and the *F*-1 scores are compared. The three-year period of 2015 – 2017 is chosen because conflict events in the previous three years was a top predictor of future events (Perry, 2013).

The study area of the analysis in Western Africa includes the states of Benin, Burkina Faso, Gambia, Ghana, Guinea, Guinea-Bissau, Ivory Coast, Liberia, Mali, Niger, Nigeria, Senegal, Sierra Leone, and Togo. Western Africa was selected to provide a diverse model that could be used as a test model for future research of a predictive model for the continent and beyond. Mauritania is excluded from the study because within the initial time frame, the state only experienced three political violence events, and the addition of Mauritania reduced the ratio of events to non-events to under one percent. This study uses a region to include diffusion and influence from neighboring states, while not including the entire continent.

The event data within the study area will be aggregated into the administrative two boundaries of each state, as opposed to a grid system. Five-by-five-minute grids based on the Global Area Reference System (GARS) grids to model the sub-national variation are not suitable

for subnational analysis because the ratio of events to non-events was about 1.3%, making the data extremely biased. The administrative two boundaries (districts) for each state are provided by the United Nations Office for the Coordination of Humanitarian Affairs (OCHA, 2019). There are 1,594 administrative two districts in the study area. 568 out of 1,594 districts experienced a political violent event from 2015 – 2017, which is an event to non-event ratio of 35.63%.

Map 1: 2012 – 2017 Training Data



5.4 Explanatory Variables

Explanatory Variables:

Ethnic Composition / Fractionization

Ethnicity is a key variable in civil conflict literature (Denny & Walter, 2014; Wolff, 2006; Fearon & Laitin, 2003; Collier & Hoeffler, 2004). Perry (2013) measured ethnic fractionalization by counting all the different ethnic groups overlapping in a territory. Weezel (2017) measured ethnic groups excluded from power. Geo-referencing of ethnic groups (GREG) data provided by Weidmann, Rød, & Cederman (2010) is used. This study will model ethnicity two different ways. First, by spatial joining the cultural group of each ethnicity into the administrative two districts; this will represent which cultural group has the largest area within the district. An advantage of using the Random Forest model is the ability to model variables by category, as opposed to relying on numeric metrics. The second method will measure the fractionization of ethnic groups within each district. This is measured by spatially joining the ethnic groups to the administrative districts. The number of different ethnic groups in the district is counted. The range is 1 – 9 with a mean fractionalization of 2.05.

Language

Language is not typically modeled with conflict prediction, but some literature demonstrates the link of language and civil conflict (Bormann, Cederman, & Vogt, 2015). The language data is from the World Language Mapping System (WLMS) Version 16 provided by (World GeoDatasets, 2015). There are too many languages to process so the languages are aggregated to language family type. Within the study area there are three language families: Afro-Asiatic, Nilo-Saharan, and Niger-Congo. The language family is aggregated to the districts by the majority of area within the district.

Diamond Deposits

The presence of diamonds is believed to play a significant role in conflict onset and duration of civil conflict in Africa and is a proxy for natural resources. Diamonds are a prominent factor in civil war literature and represent lootable resources well (Lujala, Gleditsch, & Gilmore, 2005; Addison et al. 2002; Perry, 2013). The data comes from Diamond Resources (Gilmore et al. 2005) dataset from the Peace Research Institute Oslo (PRIO). All the diamond data is considered because even if the diamond deposit requires heavy machinery, the rebel group or violent extremist group can still extort rent from the deposit. The data is geo-referenced within the study area and converted from points to polygons using a buffer around the points. The buffer is 20 km, which is roughly double the individual grid sizes. The buffered diamond deposits are intersected with the districts to create a binary variable indicating whether diamonds are located there or not. There are 147 districts that contain diamond deposits.

Petroleum Locations

The existence of petroleum is the second proxy for natural resources. Perry (2013) and Weezel (2017) both use presence of petroleum as a predictor for conflict. The data is provided by the PRIO Petroleum Dataset v. 1.2 (Lujala, Rød, & Thieme, 2007). The data consist of polygons of petroleum deposits around the world. The data is geo-referenced within the study area and intersected with the districts to create a binary variable, same as the Diamond Deposits dataset. There are 197 districts that contain petroleum deposits.

Violent Extremist Organization Presence

Districts with a violent extremist organization are more likely to experience an increased risk of political violence than districts with no violent extremist presence. Inspired by Nemeth, Mauslein, & Stapley (2014), analysis of terrorist hot spots, presence of violent extremist

organizations is modeled by spatial joining terrorist incidents from the Global Terrorism Dataset from 2015 – 2017 from the Study of Terrorism and Reponses to Terrorism (START; 2019) to the districts. 311 districts experienced the presence of a VEO group.

Previous Conflict Events

Previous conflict events are a strong predictor for future conflict events in previous machine learning models (Weezel, 2017; Perry, 2013). The ACLED political violence data from 2012 – 2014 is spatial joined to the districts and measured using a binary value of presence of previous conflict events or not. There are 551 districts that experienced a political violence event from 2012 – 2014.

Area of Districts

Territory has been linked to conflict in a number of studies (Diehl, 1991; Huth, 1996; Kalyvas, 2006). Area of the districts is used as a variable to account for conflict fought over territory. It also could be used as another proxy for areas of sparse population. Typically, larger districts are more associated with rural populations, compared to smaller districts. The difference between rural and urban areas have been shown to correlate with conflict patterns (Buhaug and Rod, 2006; Weezel, 2017). Area of districts is calculated directly within the ArcGIS software. It is ranged from 9,178 square kilometers – 332 million square kilometers, with a mean of 3 million square kilometers.

Polity IV Score

The Polity IV dataset from the Center for Systemic Peace (Marshall, Gurr, & Jaggers, 2017) is used to model governance. Perry (2013) models governance by whether the head of government if from the military and vote share of government and opposition coalitions in the legislative branch. The Polity IV dataset is used because it provides temporal quantitative scores

of the different states within the study area. The data ranges from 10, which is equal to a full democracy to -10, which is an Autocracy. This study took the average score from 2012 – 2017, which ranged from -3.5 – 8, with a mean of 5.467.

Explanatory Distance Variables:

Distance to Capital

Conflict in relation to the distance of the capital city has previously been used to describe where conflict will take place. Buhaug and Rod (2006) found conflict is more likely near the capital city in governmental conflicts but more likely away from the capital city in territorial conflicts, demonstrating the link between the distance of the capital city and conflict. Schutte (2016) used distance to the capital city as a predictor of conflict. The state capital locations are geolocated and downloaded from ESRI. The centroid of each district will be measured to the capital city by the near function, built into the Random Forest algorithm.

Distance to Border

Similar to distance from the conflict to capital city, distance to the border is an important indicator of the location of conflict events because international borders can provide refuge to rebels (Schutte, 2016; Salehyan, 2009; Buhaug, 2005). The international borders (OCHA, 2019) are provided by dissolving the districts by the state name. The result of the spatial analysis is administrative boundaries for all the states in the study area.

Distance to Natural Resources

In addition to the presence of natural resources, distance to petroleum locations and diamond locations are also modeled. The distance feature is another method to measure the importance of natural resources as predictor of political violent events. If natural resources are linked to conflict, then distance between natural resources and conflict should be an important

factor, based on the framework of Loss of Strength Gradient. The combination of presence and distance measures accurately represents the spatial aspect of natural resources.

Explanatory Rasters:

Land Cover

Densely forested areas are a common metric used in civil war literature. The causal role of land cover remains disputed but general correlations between land cover and violence is widely accepted (Schutte, 2016). Global land cover datasets have been produced but the data is infrequently updated. This study will use the global land cover data produced in 2012 (Channan, Collins, & Emanuel, 2014). The land cover types are aggregated by type to increase the predictive power; there are ten types of land cover classifications. The data is treated as a categorical type.

Economic Activity

Nighttime light activity is used as a proxy for economic development and measured by luminosity within a given grid cell. Gross Domestic Product (GDP) per capita, measuring economic output, has been linked to conflict prediction models (Muchlinski et al., 2015; Perry, 2013). Unfortunately, sub-national economic data is difficult to obtain over multiple states. Therefore, Nighttime Lights is a useful proxy in developing states where data is of poor quality; it is used in a number of conflict prediction studies (Weezel, 2017; Lazicky, 2017). Hegre et al. (2017) provide the data for nighttime light activity for 2012. The values range from 0 – 26.864, with a mean of 0.488.

Terrain and Slope

In civil war literature, terrain is usually measured as percent of mountainous terrain of the total area (Collier & Hoeffler, 2004) or in sub-national literature as the share of grid-cell

(Weezel, 2017). This study will use data from the Shuttle Radar Topography Mission (SRTM; 2014), which provides elevation data at a 90-meter resolution in a raster data format. The use of SRTM will provide a near exact elevation of each conflict incident and the random forest decision tree model will be able to provide a correlation between the conflict incidents and the elevation associated with the event without using a proxy measure, such as a percent of mountainous terrain. In addition to elevation data, slope is included to model areas of rough terrain. Slope is derived from the SRTM elevation data utilizing a tool in ArcGIS. Slope is another variable to measure mountainous areas. Terrain ranges from 0.368 feet – 1,349.677 feet, with a mean elevation of 242.014; Slope ranges from 0.08 – 14.466, with a mean slope of 2.219.

Infant Mortality Rate and Percent of Children 0 – 14

Poverty is a common socio-economic metric related to models of conflict. Infant mortality rate is one of the proxies used to measure poverty. The Global Subnational Prevalence of Child Malnutrition and the Global Subnational Infant Mortality Rates (CIESIN, 2015) provides a raster grid of the subnational infant mortality rates, measured child mortality rate per 10,000 live births. Poverty has been linked to conflict (Collier & Hoeffler, 2004; Weezel, 2017); furthermore, Muchlinski, et al. (2015) find infant mortality rate the fifth most important variable for their Random Forest model Another proxy for poverty used is percent of children aged 0 – 14 of the general population. The data is derived from the Gridded Population of the World Basic Demographic Characteristics (CIESIN, 2010). The percent is calculated by dividing the number of people aged 0 -14 per grid cell by the number of people per grid cell. The data is ranged from 26.65% – 55.84%, with a mean of 43.39%.

Population Density

Civilian population concentrations are considered a predictor of conflict events (Raleigh & Hegre, 2009; Schutte, 2016). Population is a common metric in civil war onset literature and conflict prediction models (Perry, 2013; Weezel, 2017). Insurgents also seek to use populated areas to extend their geographic control over relevant parts of the state (Kalyvas, 2006). The population density is from WorldPop (Linard et al., 2012). The data is measured by the total number of people per grid square at a one-kilometer resolution; the study will use population data from 2015. The data is ranged from 0.076 – 59,641.73, with a mean of 719.311.

Conflict Density

A raster of the density of political violent events is another method used to model previous conflict areas. Using spatial analysis, a kernel density is performed on the ACLED data and rescaled from 1 – 10. Ten is the highest concentration of political violent events and one is the least concentrated. A variable like Conflict Density has not been included in previous prediction studies. The inclusion of the variable will be monitored for data redundancy with similar variables; a valid model with the inclusion of the variable could enable future research into the link between spatial clustering and conflict prediction. The model will treat this data as a categorical type; the mean is 5.609.

Distance to Major Road Intersections and Road Density

Accessibility is usually measured by travel times between cities with more than 50,000 inhabitants (Schutte, 2016; Perry, 2013). Incorporating the literature of strategic locations (Hammond, 2018), this study uses road density and distance to major road intersections to get two perspectives of accessibility. Distance to major road intersections will model strategic locations that could be targeted. Road Density will model accessibility to the road infrastructure.

The roads data is from OpenStreetMap for all the states in the study area. Spatial analysis was preformed to create the density, where it was scaled 1 -- 10 in terms of least dense to most dense. The model will digest the data as a categorical type, with a mean of 7.88. The distance to major road intersection ranges from 0.171 km – 613.837 km, mean of 19.391 km.

5.5 Procedures

Once the data is collected, it must be geoprocessed for the two models. The analysis is aggregated into administrative two boundaries, or districts. Spatial topology is performed on the individual states to conflate the errors of overlapping and gaps between borders. Utilizing the spatial join tool in ArcMap, the conflict incidence points, ethnicity polygon, petroleum polygon, diamond deposits polygon, and language polygon were joined with the districts. This will allow for the Random Forest algorithm to analyze the data as explanatory variables. The raster data types: land cover, economic activity, terrain, slope, population density, poverty, and accessibility of road infrastructure was transformed using the resample tool in ArcMap to get the average of the values within the pixel. Without the resample, the Random Forest model will take the value at the center of the grid; that value would not represent the grid and skew the data.

To build spatial predictions, a Random Forest model will be created utilizing the Forest-based Classification and Regression tool provided by ESRI. A baseline classification model is trained using an in-sample technique of conflict incidence locations aggregated to the districts from 2015 – 2017, with 30% of the data excluded for validation. The classification model will examine a binary outcome of conflict occurrence in the districts**.** Twenty-three explanatory variables will be split across the three categories of dependent variables: explanatory training variables, explanatory training distance variables, and explanatory training rasters. Once the model is trained, the predictive power of the model is tested by utilizing an out-of-sample

technique. The accuracy of the model is tested by comparing the location of predicted conflict within the districts versus actual conflict incidence locations from 2018, using a *F*-1 score.

Next, the data is converted to format the Generalized Linear Regression. The Generalized Linear regression model does not support raster data types; therefore, the data is converted into explanatory training variables, but the distance variables remain. The model is trained and predicted to assess against the out-of-sample 2018 data. Once the model is complete, the regression residuals are assessed for spatial autocorrelation by running the Moran's I statistical test. The residuals should be spatial random for a reliable model. An *F*-1 score is then calculated and compared to the Random Forest model to assess the predictive performance of the two models.

Chapter Six: Results

In order to compare the predictive power, two models were applied: Classification Random Forest Model and the Logistic Generalized Linear Regression model. Following machine learning convention, the data was randomly divided with 70 percent of the observations allocated to the training set and 30 percent for the test set for the in-sample predictions from 2015 – 2017. The two models were tested against out of sample conflict incidence data from 2018 and compared using a *F*-1 score.

6.1 Model Performance

Out of Bag Errors are the primary way to test the model performance of the Random Forest model. The model provides error scores for half the trees and the total number of trees to evaluate if increasing the total number of trees improves the performance. The model tested up to 250 trees, but the model did not improve after 150 trees. The Mean Squared Error (MSE) score was 27.551 at the 150^{th} tree. The percent of variance explained is split by classification type. In districts where conflict occurrence was not predicted, the model only accounted for 16.925% of the variation explained; compared to 45.833% in districts of predicted conflict occurrence. This result is unsurprising since the model overpredicted areas of conflict incidence, as opposed to underpredicting areas of non-conflict incidences.

The Generalized Linear Regression model has a number of different metrics to assess the validity and statistical significance of the model. The Joint Wald Statistic, which indicates the overall model significance, displays a p-value of less than 0.01. In other words, the overall model is statistically significant. The percent of deviance explained is the proportion of dependent variable variance accounted for by the explanatory variable. The explanatory variables only explained 29 percent of the variance, which is low. This indicates explanatory variables are

missing, which could improve the model performance. The Variance Inflation Factor (VIF) indicates redundancy among explanatory variables; values over 7.5 are considered redundant. None of the variables included violated this assumption. This last check of the model performance is to test the deviance residuals for spatial autocorrelation. When there is statistically significant spatial autocorrelation of the regression residuals, the Generalized Linear Regression model will be considered incorrectly specified and, consequently, results from Generalized Linear Regression are unreliable. The residuals are spatial random; therefore the model is considered reliable.

Figure 2: Accuracy of True Predictions

	Accuracy
Random Forest	74.72 %
Generalized Linear Regression	78.356 %

Figure 2 displays the accuracy performance of the algorithms, without accounting for incorrect predictions. The Generalized Linear Regression model displayed better accuracy, yet there is not a large variation between the two models in terms of accuracy. According to Perry (2013), this is not surprising as prior research indicates that machine learning tends to over predict instances of conflict and districts experiencing conflict occurrence are rare; therefore, large swings in accuracy would have little effect on the overall accuracy. Of the 1,594 districts within the study area, the Random Forest correctly predicted 1,191. The Generalized Linear Regression model correctly predicted 1,249 of the 1,594 districts.

6.2 *F*-1 Score Results

Overall, the two models *F*-1 Scores are remarkably similar. The Generalized Linear Regression model slightly outperformed the Random Forest model, with a *F*-1 score of 0.61017 compared to 0.58582, although the difference is not significant. The Random Forest Model

predicted conflict occurrence in areas of actual conflict occurrence in the out of sample data better than the Generalized Linear Regression model, but not by a wide margin. In contrast, the Generalized Linear Regression model predicted areas of non-conflict better than the Random Forest model by a significant margin. The most significant difference between the two models is the overprediction of conflict in areas where conflict occurrence was not present in the 2018 out-of-sample data by the Random Forest model. The Random Forest model overpredicted 73 more occurrences than the Generalized Linear Regression model, lowering the Precision score of the Random Forest model. The Random Forest model predicted more true values correctly than the Generalized Linear Regression model, but the overpredicting caused its Precision score to fall below the Generalized Linear Regression model. The Generalized Linear Regression model underpredicted areas of actual conflict occurrence more than the Random Forest model, which results in the Random Forest possessing a better Sensitivity score.

Figure 3: Random Forest Confusion Matrix

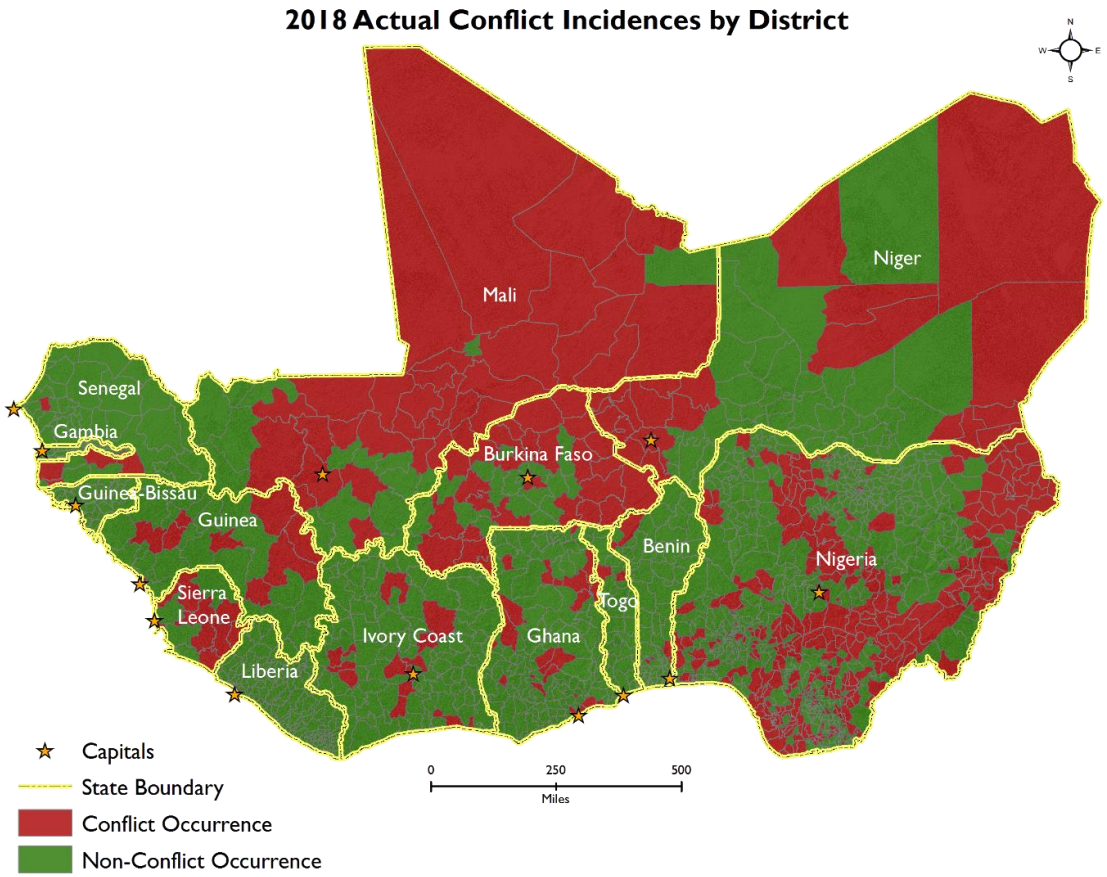
	Predicted Class		
Actual Class		Conflict = 1	Conflict = 0
	Conflict = 1	285	159
	Conflict = 0	244	906
Precision	0.53875		
Sensitivity	0.64189		
F-1 Score	0.58582		

Figure 4: Generalized Linear Regression Confusion Matrix

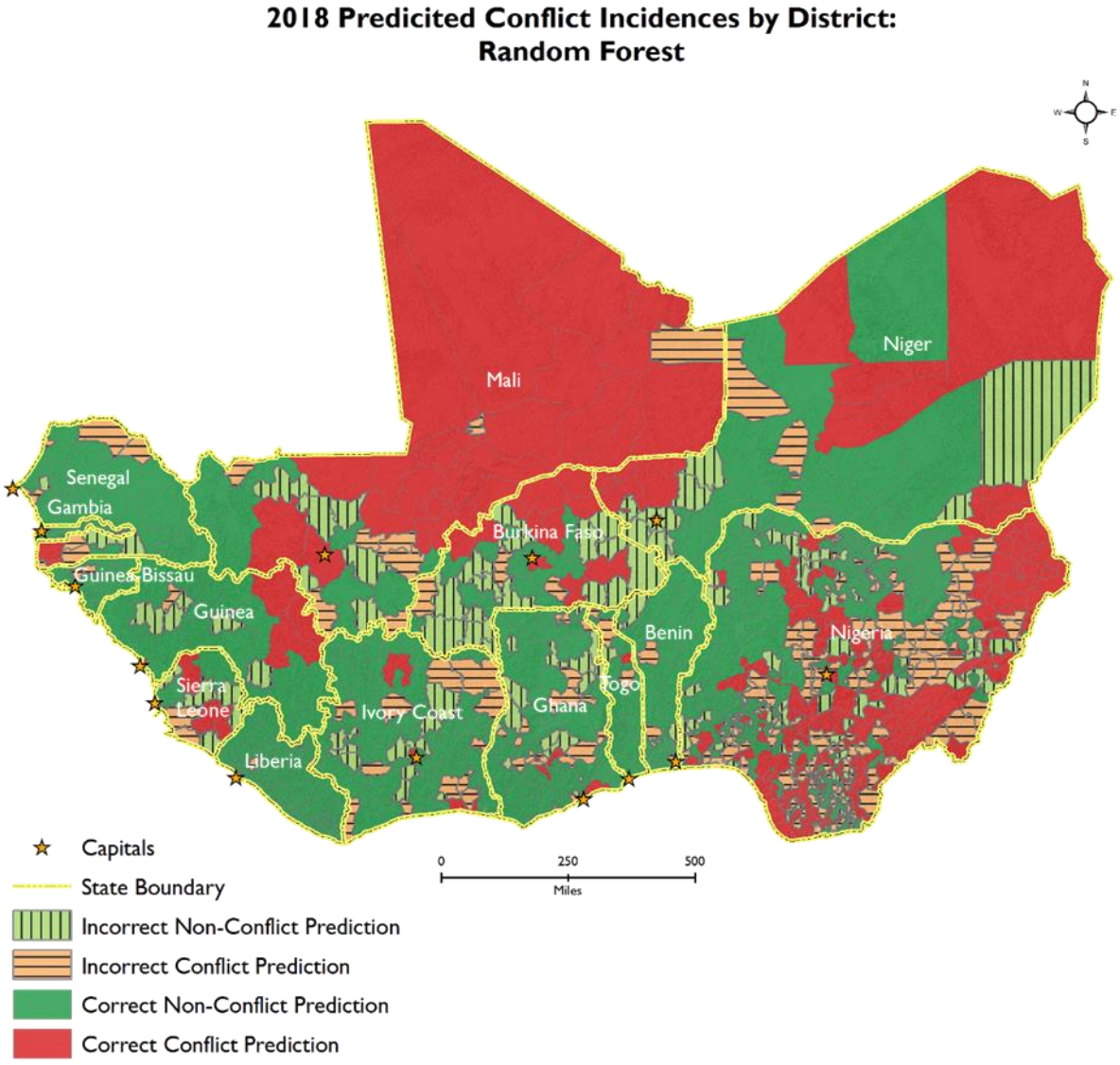
	Predicted Class		
Actual Class		Conflict = 1	Conflict = 0
	Conflict = 1	270	174
	Conflict = 0	171	979
Precision	0.61224		
Sensitivity	0.60811		
F-1 Score	0.61017		

A common critique of machine learning model results is the ability to communicate the results clearly for decision makers to analyze the data (Cederman & Weidmann, 2017); maps are an effective tool to display the predicted results and Confusion Matrix. Map 2 displays the out-of-sample 2018 ACLED conflict incidence data; this represents the actual class. Overall, both maps display large areas of consideration, in regards to conflict occurrence, primarily in Mali, Niger, and Nigeria.

Map 2: 2018 Out-of-Sample True Data

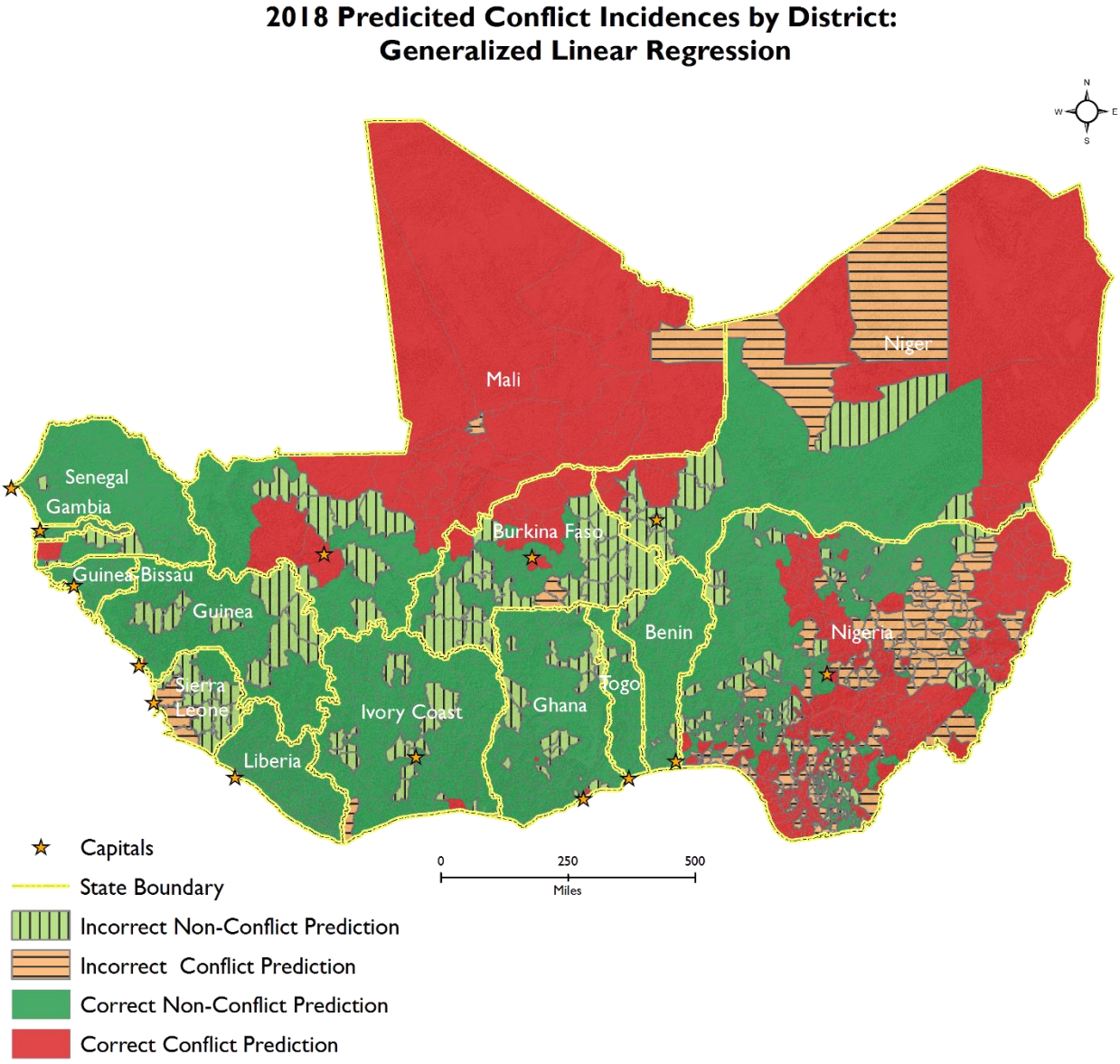


Map 3: Random Forest Conflict Prediction Map



Spatially, the Random Forest model performed well predicting conflict occurrence in Mali and capturing some level of conflict occurrence in the states along the coast. Nigeria is fairly overpredicted, especially in the northeast region. Within Burkina Faso, Ivory Coast, Ghana, and Togo, the model appears to be spatially close in accurately predicting districts of conflict occurrence. This is evident by predicted districts of conflict occurrence neighboring districts with predicted non-conflict occurrence.

Map 4: Generalized Linear Regression Conflict Prediction Map



Spatially, the Generalized Linear Regression model predicted conflict occurrence along the outer edge of the study area. Similar to the Random Forest model, conflict occurrence was overpredicted in Northeast Nigeria. The Generalized Linear Regression model primarily underpredicted conflict incidences in the interior of the study area, in Ghana, Togo, Burkina Faso, Ivory Coast, and Guinea.

6.3 Explanatory Variables Results

Figure 5 displays the Random Forest model list of the top 20 variables, listed by importance, or the number of times a variable is responsible for a split and the impact of that split divided by the number of trees. Overall, the geographic explanatory variables were more important in the Random Forest model than the Generalized Linear Regression model. Land Cover and Slope are important in the Random Forest model but not in the Generalized Linear Regression model. Distance explanatory variables, Distance to State Boundaries and State Capitals were important in the random Forest Model but not significant in the GLR model. The top variable is Cultural Group of the different ethnicities. An advantage of the Random Forest model over other statistical models is the ability to utilize non-numerical data, therefore the Generalized Linear Regression model is not able to use Cultural Group as a variable. Conflict Density and Road Density are the next two important variables for the Random Forest Model. Both variables were created utilizing spatial analysis to create the data. Because conflict occurrence is rare, conflict density acts as an area reduction variable. Areas of high density of conflict are more likely to experience conflict occurrence in the future.

Figure 5: Random Forest Top Variable Importance

Random Forest Top Variable Importance		
Variable	Importance	%
CULTUREGRP	3.09	8
CONFLICT DENSITY	2.63	7
ROAD DENSITY	2.52	6
LAND COVER	2.09	5
POPULATION DENSITY	2.08	5
SLOPE	2.01	5
AREA OF DISTRICT	2.00	5
DIST TO ADMIN	2.00	5
DIST TO CAPITAL	1.94	5
NIGHTTIME LIGHTS	1.93	5
TERRAIN	1.92	5

PERCENT CHILDREN	1.89	5
INFANT MORT RATE	1.87	5
DIST ROAD INTSEC	1.83	5
DIAMOND DIST	1.81	5
OIL DIST	1.75	4
ETHNIC FRAC	1.74	4
POLITY IV SCORE	1.31	3
VEO PRESENCE	1.05	3
PREVIOUS INCIDENTS	0.91	2

Figure 6 presents the explanatory variables and displays the probability of the variables of statistical significance to the Generalized Linear Regression model. The three most statistically significant explanatory variables are Previous Onset, Violent Extremist Organization (VEO) onset, and Conflict Density. All three variables signify areas of a high probability of conflict occurrence. Districts that have the presence of VEOs and previous conflict are more likely to experience conflict in the future. Unexpectedly, in the Random Forest model, previous conflict occurrence from 2012 – 2014 did not contribute to the overall performance of the model. Perry (2013) found previous battles to be the number one contributing variable to his model. Previous conflict occurrence did contribute to the Generalized Linear Regression model, however.

Figure 6: Summary of Generalized Linear Regression Results

Summary of GLR Results					
Variable	Coefficient	StdError	Z-Statistic	Probability	VIF
Intercept	-2.069811	1.154087	-1.793462	0.072899	----------
POLITY IV SCORE**	-0.046275	0.052239	-0.885831	0.375709	1.163666
AREA OF DISTRICT	0.000000	0.000000	2.356839	0.018431*	2.877497
PREVIOUS INCIDENTS	0.823881	0.142364	5.787150	0.000000*	1.345237
OIL PRESENCE	0.052194	0.252881	0.206399	0.836479	1.842253
DIAMOND PRESENCE	0.263577	0.262102	1.005627	0.314595	1.332635
ETHNIC FRAC	0.190310	0.062531	3.043445	0.002339*	1.290367
DIST TO CAPITAL	-0.000327	0.000607	-0.537918	0.590634	2.074883
DIST TO ADMIN	-0.000232	0.001089	-0.213449	0.830977	1.782652
VEO PRESENCE	1.196321	0.189032	6.328661	0.000000*	1.572408
DIAMOND DIST	-0.000000	0.000001	-0.602884	0.546586	3.795079

OIL DIST	0.000001	0.000000	2.915215	0.003554*	2.706231
TERRAIN	-0.000656	0.000569	-1.152323	0.249188	2.594962
SLOPE	0.004703	0.051721	0.090932	0.090932	1.732867
ROAD DENSITY	-0.135124	0.039326	-3.435980	0.000590*	1.841714
PERCENT CHILDREN	-1.780134	2.173828	-0.818894	0.412847	2.570672
DIST ROAD INTSEC	0.000142	0.003962	0.035956	0.971318	3.409186
CONFLICT DENSITY	0.394249	0.047672	8.269970	0.000000*	3.632487
POPULATION DENSITY	-0.000077	0.000032	-2.431488	0.015037*	2.071396
NIGHTTIME LIGHTS	0.162535	0.058175	2.793894	0.005208*	2.208781
LANDCOVER	0.018882	0.027255	0.692806	0.488432	1.263216
INFANT MORT RATE	0.000077	0.000128	0.603078	0.546457	1.139493

Of the nine statistically significant variables for the Generalized Linear Regression model, five explanatory variables are statistically significant in the Generalized Linear Regression model and a top variable of importance in the Random Forest model. The five variables are: area of the districts, road density, conflict density, population density, and nighttime lights. Four of the five overlapping variables utilized spatial analysis to create and/or process the data and are raster feature types. Nighttime lights, population density, and road density tend to highlight urbanized districts, as high economic activity, road infrastructure, and population are the characteristics of a city, as opposed to a rural area. Conflict density acts as an area reduction tool to highlight areas of clustering, though not statistically significant cluster in the form of Hot Spot Analysis. The outlier of the five explanatory variables is the size of the district. The size of the districts varies greatly from the rural districts in some states, such as Mali, to some districts in state's which divide metropolitan areas into separate districts.

Chapter Seven: Discussion

Conflict prediction models in quantitative international relations literature have incorporated machine learning methods and spatial analysis methods into prediction models, but rarely do prediction models incorporate both aspects. In this paper, two types of spatial models were compared to assess which model possessed the superior predictive power. A Classification Random Forest model and Logistic Generalized Linear Regression model were separately trained and tested using conflict data from 2015 – 2017 against 23 explanatory variables, including distance and raster features, to provide an in-sample test. The models then predicted conflict occurrence for 2018 and testing against 2018 out of sample data. The Generalized Linear Regression model produced a *F*-1 score of 0.61017, compared to a *F*-1 score of 0.58582 for the Random Forest model. While the scores are similar, the Generalized Linear Regression model is overall more accurate, in part because of the Random Forest model's overprediction of conflict occurrence when no conflict was present. Five variables stuck out as important for both models: Conflict Density, Road Density, Nighttime Lights, Population Density, and Area of the District.

Previous conflict prediction literature indicated Random Forest models have superior predictive power over a more traditional logistic model. While the Random Forest model predicted more correct incidence of conflict occurrence than the Generalized Linear Regression model, the overall performance of the Generalized Linear Regression was better than the Random Forest model, albeit not by a wide margin. This is significant because a Generalized Linear Regression model is rarely utilized in the conflict prediction literature. Although, both methods show promising capabilities for conflict prediction. The Random Forest model is similar to the overall accuracy (False versus true positives) score achieved by Perry (2013), which produced an accuracy score of 58.5% for the Random Forest and 24.6% for the Naïve

Bayes model. The Generalized Linear Regression model is an improvement over all three models. Muchlinski et al. (2015) compared Random Forest models with Logistic Regression models to predict civil war onset and all the logistic models failed to predict onset of civil wars in the out-of-sample data, while the Random Forests correctly predicted nine of the 20 civil war onsets in the out-of-sample data. Their results are contrary to my findings, was the logistic regression model outperformed the Random Forest when applied to conflict occurrence. A key difference between the two studies is the lack of use on the Logistic Generalized Regression model. Furthermore, Weezel (2017) produced a subnational conflict logistic model in Africa with geographic and socio-economic explanatory variables; his model has a *F*-1 score of 0.257. By comparison, the Random Forest model produced a *F*-1 score that nearly doubled at 0.58582 and the Generalized Linear Regression model had a *F*-1 score of 0.60811. Wezzel (2017) model only had 11 explanatory variables, which could impact the resulting *F*-1 score. More traditional statistical models, like the Generalized Linear Regression should merit consideration for future conflict prediction models, as the focus has shifted toward machine learning approaches.

Five explanatory variables were significant to both models analyzed: Conflict Density, Road Density, Nighttime Lights, Population Density, and Area of the District. Three of the five explanatory variables are spatial densities. A positive attribute of the Random Forest model is the ability to process raster data. Raster data is an integral aspect of spatial analysis and the bridge between spatial analysis and prediction models; Nighttime Lights is also a raster dataset. The densities display concentrations of a spatial phenonium and can act as an area reduction method for the prediction models. Road Density and Conflict Density have never been used as a variable in the previous research. In the Generalized Linear Regression model, Nighttime Lights displayed significant explanatory power; this finding is contrary to Weezel (2017), which found

little explanatory power in the Nighttime Lights variable. Previous Conflict occurrence from 2012 – 2015 was statistically significant in Generalized Linear Regression model but surprisingly not the Random Forest model. Perry's (2013) top variable importance was previous battles spanning three years. Surprisingly, area of the district is a variable important to both models. Typically, larger districts indicate more rural areas. Conflict is generally more likely closer to urbanized areas (Weezel, 2017; Buhaug & Rod, 2006). Area of districts could be a new variable to measure this theory. The presence and distance of natural resources had little effect on the explanatory power of the two models, though distance to oil fields was statistically significant in the Generalized Linear Regression model. Weezel (2017) and Perry (2013) also found natural resources contributing little to their models.

The Conflict Density explanatory variable produced fascinating results for both models. It was statistically significant in the Generalized Linear Regression model and the second most important variable. A variable such as a spatial density of the conflict locations has never been added as an explanatory variable in a prediction model. Data redundancy was a concern when added to both models, but the variable ended up playing a critical role in the models. The clustering of conflict incidences has been studied in previous research (Braithwaite, 2010a; Lujala, 2009), but a link between conflict hot spots and conflict prediction has yet to be explored.

Better sub-national data is needed in order to improve the model performance and increase the percent of variance explained, especially geospatial focused data. Perry (2013) came to the same conclusion after his initial feasibility study into Random Forest modelling of conflict. While the Generalized Linear Regression model was statistically significant, the exploratory variables only explained 29 percent of the variance and the Random Forest model

only explained about 31 percent of the variance but did explain 46 percent of the variance for predicted conflict locations.

It is possible with more refined spatial data at the sub-national level of analysis, the results of the two models could change. Sub-national data depicting education, such as secondary education levels and food insecurities could help refine the models. Better sub-national economic data could improve model results, instead of relying on proxies, such as Nighttime Lights. The error score for the Random Forest was a bit high; further refinement of the model could improve results.

This study is not without a few limitations. When dealing with spatial data, issues arise when selecting a study area because of the modifiable areal unit problem associated with spatial data. The results will vary between unit of analysis, for example examining conflict occurrence in a single state will have drastic result differences compared to a continent, such as Africa. A regional approach was chosen to capture diffusion of conflict but also maintaining a more local model. The larger the scale of the model, the more difficult it becomes to find suitable spatial-temporal data to cover the study area. Geospatial data aggregation is crucial in spatial analysis. The explanatory variables were aggregated to the district level of analysis in this study, which can vary in size and change at the state's discretion. The raster explanatory variables would provide more value in a grid-based unit of analysis. A grid-based unit of analysis would unlock the true geographic values of variables such as Terrain, with relying on averages within a district. Unfortunately, a grid-based system creates bias ratios of events to non-events. Future modeling techniques can help overcome the bias ratio to unlock the value of geographic data.

The performance of the Generalized Linear Regression model should warrant future consideration as a viable option for conflict prediction. The Generalized Linear Regression

model is considered a global model, meaning it considers the area of study as a single unit, without considering the local variation. The next logical step for future research is to compare the Generalized Linear Regression, with a local model, such as a Logistic Geographic Weighted Regression model. The models can be easily compared using the AICc metric. Furthermore, the performance of the Conflict Density variable and the effect of conflict hot spots can be explored further. Hot spot analysis is a local measure of spatial autocorrelation. Conflict prediction research could explore the local variation of conflict incidences and its relation to conflict prediction. Lastly, as geospatial data becomes more readily available, especially grid-based geospatial data, the literature on conflict prediction would greatly benefit from studies addressing the scale of the study area. Future research analyzing the effects of the modifiable areal unit problem of a grid-based system versus a more traditional administrative boundary system would help researchers better understand the spatial aspect of conflict, which could lead to better predictions.

Conclusion

This study compared the predictive performance of two spatial models for conflict occurrence in Western Africa: a Classification Random Forest model and a Logistic Generalized Linear Regression model. Surprisingly, the Generalized Linear Regression model produced a more accurate model, with a *F*-1 score of 0.61017, than the Random Forest model, with a *F*-1 score of 0.58582. While the Generalized Linear Regression model performed better, the difference is not significant enough to declare one method superior to the other. The Random Forest model predicted incidence of true conflict occurrence better than the Generalized Linear Regression model. Of the twenty-three explanatory variables, five variables contributed to the explanatory power of both models: Area of the Districts, Road Density, Conflict Density, Population Density, and Nighttime Lights. Conflict Density provides an opportunity for future research to link the association of conflict hot spots with conflict prediction. Spatial-temporal hot spots of conflict incidences could contribute to the overall variance explained, while improving the prediction model.

A spatial approach, with the incorporation of spatial data, demonstrates promise for future research. The further refinement of spatial data at the subnational level could produce better results. The Generalized Linear Regression *F*-1 score opens up possibilities to further explore this method at different scales of analysis. Since the Generalized Linear Regression model is a global model, the next logical step is to incorporate a local model, the Geographic Weighted Regression, to further explore the local variation of conflict occurrence. A local model can be compared to the global model, which will assess the benefit of exploring the local variation. Combined with a spatial-temporal hot spot analysis, exploring the local variation of conflict incidences could help further the existing body of literature.

Bibliography

Addison, T., Billon, P. L., & Murshed, S. M. (2002). Conflict in Africa: The Cost of Peaceful Behavior. *Journal of African Economies*, 11(3), 365 – 386.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht, Netherlands: Kluwer.

Anselin, L. (1995). Local Indicators of Spatial Association - LISA. *Geographical Analysis*, 27(1), 93 – 115.

Anselin, L., Florax G.M. & Rey, S.J. (eds) (2004) Advances in Spatial Econometrics: Methodology, Tools, and Applications. New York: Springer.

Anselin, L., & O'Loughlin, J. (1990). Spatial Econometric Analysis of International Conflict. In Chatterji, M. & Kuenne, R. E. (eds), *Dynamics and Conflict in Regional Structural Change*. New York: New York University Press, 325 – 345.

Anselin, L., & O'Loughlin, J. (1992). Geography of International Conflict and Cooperation: Spatial Dependence and Regional Conflict in Africa. In Ward, M. D. (eds), *The New Geopolitics*. Philadelphia: Gordon and Breach, 39 – 75.

Auty, R. (2004). Natural Resources and Civil Strife: A Two-Stage Process. *Geopolitics*, 9:1, 29 – 49.

Bates, R. (1987). *Essays on the Political Economy of Rural Africa*. Berkeley: University of California Press.

Basuchoudhary, A., Bang, J., Sen, T., & David, J. (2018). Using Machine Learning To Predict Conflict: Toward an unified theory of civil conflict? Working Paper.

Beck, N., King, G., & Zeng, L. (2000). *American Political Science Review*, 94, 21 – 35.

Berdal, M. & Malone, D. M. (2000). *Greed & Grievance: Economic Agendas in Civil Wars*. Boulder, CO: Lynne Rienner

Blair, R., Blattman, C., Hartman, A. (2017). Predicating Local Violence: Evidence from a Panel Survey in Liberia. *Journal of Peace Research*, 54, 298 – 312.

Blair, R. & Sambanis, N. (2016). Forecasting Civil Wars. Working paper.

Blattman, C. & Miguel, E. (2010). Civil War. *Journal of Economic Literature*, 48(1), 3 -57.

Bond, J., Petroff, V., O'Brein, S., Bond, S. (2004). Forecasting Turmoil in Indonesia: An Application of Hidden Markov Models. Presented at the International Studies Association Meeting, Montreal.

Bormann, N., Cederman, L., & Vogt, M. (2015). Language, Religion, and Ethnic Civil War. *Journal of Conflict Resolution*, 61(4), 744 – 771.

Boulding, K. (1962). *Conflict and Defense*. New York: Harper and Row.

Braithwaite, A. (2006). The Geographic Spread of Militarized Disputes. *Journal of Peace Research*, 43 (5), 507 – 522.

Braithwaite, A. (2010a). *Conflict Hot Spots: Emergence, Causes, and Consequences*. Farnham, England: Ashgate.

Braithwaite, A. (2010b). MIDLOC: Introducing the Militarized Interstate Dispute Location Dataset. *Journal of Peace Research*, 47(1), 91 – 98.

Branch, J. (2016). Geographic Information Systems (GIS) in International Relations. *International Organization*, 70, 845 – 869.

Brandt, P. T. & Freeman, J. R. (2005). Advances in Bayesian Time Series Modeling and the Study of Politics: Theory Testing, Forecasting, and Policy Analysis. *Political Analysis*, 14(1), 1 – 36.

Brandt, P. T., Colaresi, M., & Freeman, J. R. (2008). The Dynamics of Reciprocity, Accountability, and Credibility. *Journal of Conflict Resolution*, 52(3), 343 – 374.

Brandt, P. T., Freeman, J. R., & Schrodt, P. A. (2014). Evaluating Forecast of Political Conflict Dynamics. *International Journal of Forecasting*, 30(4), 944 – 962.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5 – 32.

Bremer, S. A. (1982). The Contagiousness of Coercion: The Spread of Serious International Disputes 1900-1976. *International Interactions*, 9(1), 29 – 55.

Bremer, S.A. (1992). Dangerous Dyads. *Journal of Conflict Resolution*, 36(2), 309 – 341.

Bueno de Mesquita, B. (1981). *The War Trap*. New Haven: Yale University Press.

Buhaug, H. (2005). *The Geography of Armed Civil Conflict*. (Unpublished Doctoral Dissertation). Norwegian University of Science and Technology, Trondheim.

Buhaug, H. & Gates, S. (2002). The Geography of Civil War. *Journal of Peace Research*, 39(4), 417 – 433.

Buhaug, H. & Gleditsch, K. S. (2008). Contagion or Confusion? Why Conflicts Cluster in Space. *International Studies Quarterly*, 52(2), 215 – 233.

Buhaug, H. & Lujala, P. (2005). Accounting for Scale: Measuring Geography in Quantitative Studies of Civil War. *Political Geography*, 24(4), 399 – 418.

Buhaug, H., & Rød, J. K. (2006). Local determinants of African Civil Wars, 1970–2001. *Political Geography*, 25 (3), 315–335.

Buhaug, H., Gates, S., & Lujala, P. (2009). Geography, Rebel Capability, and the Duration of Civil Conflict. *Journal of Conflict Resolution*, 53(4), 544 – 569.

Cederman, L. E., & Weidmann, N. B., (2017). Predicting Armed Conflict: Time to Adjust our Expectations? *Science*, 335, 474 – 476.

Center for International Earth Science Information Network (CIESIN) – Columbia University. (2015). Basic Demographic Characteristics, v4.11 [Data file]. Retrieved from http://sedac.ciesin.columbia.edu/data/sets/gpw-v4-basic-demographic-characteristics-rev11.

Center for International Earth Science Information Network (CIESIN) – Columbia University. (2015). Global Subnational Infant Mortality Rates, v2 [Data file]. Retrieved from http://sedac.ciesin.columbia.edu/data/sets/povmap-global-subnational-infant-mortailty-rates-v2.

Celiku, B. & Kraay, A. (2017). Predicted Conflict. World Bank Policy Research Working Paper 8075.

Chadefaux, T. (2017). Conflict Forecasting and its Limits. *Data Science,* 1, 7 – 17.

Channan, S. K., Collins, K., Emanuel, W. R. (2014). Global mosaics of the standard MODIS land cover type data. University of Maryland, College Park, MD.

Cliff, A.D. & Ord, J.K. (1973). *Spatial Autocorrelation*. London: Pion.

Cliff, A.D. & Ord, J.K. (1981). *Spatial Processes: Models and Applications*. London: Pion.

Cohan, S. B. (2015). *Geopolitics: The Geography of International Relations*. Lanham, Maryland: Rowman & Littlefield.

Collier, P. & Hoeffler, A. (2004). Greed and Grievance in Civil Wars. *Oxford Economic Papers*, 56(4), 663 – 695.

Collier, P., Hoeffler, A., & Söderbom, M. (2004). On the Duration of Civil War. *Journal of Peace Research*, 41(4), 253 – 273.

Costalli, S. & Moro, F. N. (2011). The Patterns of Ethnic Settlement and Violence: A Local-L Level Quantitative Analysis of the Bosnian War. *Ethnic and Racial Studies*, 34(12), 2096 – 2114.

Darmofal, D. S. (2006). *Spatial Econometrics and Political Science*. Presented at the Annual Meeting of the Southern Political Science Association, Atlanta.

De Juan, A. (2012). Mapping Political Violence – The Approaches and Conceptual Challenges of Subnational Geospatial Analyses of Intrastate Conflict. *GIGA Working Paper*, 211.

de Soysa, I., (2000). The Resource Curse: Are Wars Driven by Rapacity or Paucity?. In M. Berdal & D. Malone, eds, *Greed and Grievance: Economic Agendas in Civil Wars*. Boulder, CO: Lynne Rienner (113–135).

DeRouen, K. R. & Sobek, D. (2004). The Dynamics of Civil War Duration and Outcome. *Journal of Peace Research*, 41(3), 303 – 320.

Denny, E. K., & Walter, B. F. (2014). Ethnicity and Civil War. *Journal of Peace Research*, 51(2), 199 – 212.

Diehl, P. (1991). Geography and War: A Review and Assessment of the Empirical Literature. *International Interactions*, 17, 11-27.

Do, Q., & Iyer, L. (2010). Geography, Poverty, and Conflict in Nepal. *Journal of Peace Research*, 47(6), 735 – 748.

Elbadawi, I. & Sambanis, N. (2002). How much War Will we see?: Explaining the Prevalence of Civil War. *Journal of Conflict Resolution*, 46(3), 307 – 334.

Ellingsen, T. (2000). Colorful Community or Ethnic Witches' Brew? Multiethnicity and D Domestic Conflict During and After the Cold War. *Journal of Conflict Resolution*, 44(2), 228 –249.

Environmental Systems Research Institute (ESRI). (2019). An overview of Modeling Spatial Relationships toolset. Retrieved from https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/an-overview-of-the-modeling-spatial-relationships-toolset.htm.

Esteban, J., & Ray, D. (2008). On the Salience of Ethnic Conflict. *American Economic Review*, 98(5), 2185 – 2202.

Faber, J. Houweling, H. W., Siccama, J. G. (1984). Diffusion of War: Some Theoretical Considerations and Empirical Evidence. *Journal of Peace Research*, 21(3), 277 – 288.

Fearon, J. D. (2004). Why do Some Civil Wars Last So Much Longer Than Others? *Journal of Peace Research*, 41 (3), 275 – 301.

Fearon, J. D. (2006). Ethnic Mobilization and Ethnic Violence. In: Weingast, B. R. & Wittman, D. A. (eds). *The Oxford Handbook of Political Economy*. New York: Oxford University Press, 852–868.

Fearon, J. D., & Laitin, D. D. (2000). Violence and the Social Construction of Ethnic Identity. *International Organization*, 54 (4), 845 – 877.

Fearon, J. & Laitin, D. (2003). Ethnicity, Insurgency, and Civil War. *American Political Science Review*, 97(1), 75 – 90.

Flint, C., Diehl, P., Scheffran, J., Vasquez, J., & Chi, S. (2009). Conceptualizing ConflictSpace: Toward a Geography of Relational Power and Embeddedness in the Analysis of Interstate Conflict. *Annals of the American Geographers*, 99(5), 827 – 835.

Fotheringham, A.S. & Wong, D.W.S. (1991) The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. *Environment & Planning A*, 23(7), 1025 – 1044.

Getis, A., & Ord, J. K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, 24(3), 189 – 206.

Gilmore, E., Gleditsch, N. P., Lujala, P., & Rød, J. K. (2005). Conflict Diamonds: A New Dataset. Conflict Management and Peace Studies, 22(3), 257 – 292.

Gleditsch, K.S. (2002). *All International Politics is Local: The Diffusion of Conflict, Integration, and Democratization*. Ann Arbor, MI: University of Michigan Press.

Gleditsch, K.S. (2007). Transnational Dimensions of Civil War. *Journal of Peace Research*, 44(3), 293 – 309.

Gleditsch, N. P. (1995). Geography, Democracy, and Peace. *International Interactions*, 4, 297 – 323.

Gleditsch, N. P. & Singer, J. D. (1975). Distance and International War, 1816 – 1965. In Proceedings of the International Peace Research Association, 481 – 506. Oslo, Norway.

Gleditsch, N.P., Wallensteen, P., Eriksson, M., Sollenberg, M., & Strand, H. (2002). Armed Conflict 1946–2001: A New Dataset. *Journal of Peace Research*, 39(5), 615 – 637.

Goldstein, J. S. (1992). A Conflict-Cooperation Scale for WEIS Events Data. *The Journal of Conflict Resolution*, 36(2), 369 – 385.

Gurr, T. R. (1970). *Why Men Rebel*. Boulder, CO: Paradigm.

Hammond, J. (2018). Maps of Mayhem: Strategic Location and Deadly Violence in Civil War. *Journal of Peace Research*, 55(1), 32 – 46.

Hegre, H. Ellingsen, T., Gates, S., & Gleditsch, N. P. (2001). Toward a Democratic Civil Peace? Democracy, Political Change, and Civil War, 1816-1992. *The American Political Science Review,* 95(1), 33 – 48.

Hegre, H., Metternich, N. W., Nyg**a**rd, H. M., & Wucherpfennig, J. (2017). Introduction: Forecasting in Peace Research. *Journal of Peace Research,* 54(2), 113 – 124.

Hegre, H., Metternich, N. W., Nyg**a**rd, H. M., Wucherpfennig, J., Weidmann, N. B., Schutte, S. (2017). Using night light emissions for the prediction of local wealth. March 2017: Special issue on Forecasting in Peace Research, 54(2), 125 – 140.

Hegre, H. & Raleigh, C. (2009). Population Size, Concentration and Civil War. *Political Geography*, 28(4), 224 – 238.

Hegre, H., & Sambanis, N. (2006). Sensitivity Analysis of Empirical Results on Civil War Onset. *Journal of Conflict Resolution*, 50 (4), 508 – 535.

Herbst, J. (2000). *States and Power in Africa. Princeton*, NJ: Princeton University Press.

Hill, D. W. & Jones, Z. M. (2014). An Empirical Evaluation of Explanations for State Repression. *American Political Science Review*, 108(3), 661 – 687.

Houweling, H. W. & Siccama, J. G. (1985). The Epidemiology of War, 1816-1980. *Journal of Conflict Resolution*, 29(4), 641 – 663.

Houweling, H. W. & Siccama, J. G. (1988). *Studies of War*. Dordrecht: Martinus Nijhoff Publishers.

Huth, P. K. (1996). *Standing Your Ground: Territorial Disputes and International Conflict*. Ann Arbor: University of Michigan Press.

Isard, W. (1954). Location Theory and Trade Theory: Short-Run Analysis. *Quarterly Journal of Economics*. 68(2), 305 – 320.

Jones, Z. & Linder, F. (2015). Proceedings from MPSA ’15: Exploratory Data Analysis Using Random Forest.

Joshi, R. (2016). Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures. Retrieved from https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/.

Kalyvas, S. N. (2006). *The Logic of Violence in Civil War*. Cambridge: Cambridge University Press.

Khatiwada, L. K. (2014). A Spatial Approach in Locating and Explaining Conflict Hot Spots in Nepal. *Eurasian Geography and Economics*, 55(2), 201 – 2017.

King, G. & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, 9(2), 137 – 163.

Koch, H., North, R., Zinnes, D. (1960). Some Theoretical Notes on Geography and International Conflict. *Journal of Conflict Resolution,* 4, 4 – 14.

Kirby, A.M & Ward, M.D. (1987). The Spatial Analysis of War and Peace. *Comparative Political Studies,* 20(3), 293 – 313.

Kjellen, R. (1916). *Staten som Lifsform*. Leipzig: Hirzel.

Lagazio, M & Russett, B. (2003). *A Neural Network Analysis of Militarized Disputes, 1885-1992: Temporal Stability and Causal Complexity*, Ann Arbor: University of Michigan Press.

Lazicky, C. (2017). *Improving Conflict Early Warning Systems for United Nations Peacekeeping* (Unpublished doctoral dissertation). Harvard University, Boston.

Leiter, H., Sheppard, E., & Sziarto, K. M. (2008). The Spatialities of Contentious Politics. *Transactions of Institute of British Geographers*, 33(2), 157 – 172.

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., & Tatem A. J. (2012). Population Distribution, Settlement Patterns and Accessibility across Africa. *PLoS One*, 7(2): e31743.

Lujala, P. (2009). Deadly Combat over Natural Resources Gems, Petroleum, Drugs, and the Severity of Armed Civil Conflict. Journal of Conflict Resolution, 53(1), 50 – 71.

Lujala, P., Gleditsch, N. P., & Gilmore, E. (2005). A Diamond Curse? Civil War and a Lootable Resource. *Journal of Conflict Resolution*, 49(4), 538 – 562.

Lujala, P., Rød, J. K., & Thieme, N. (2007). Fighting Over Oil: Introducing A New Dataset. *Conflict Management and Peace Science*, 24(3), 239 – 256.

Mackinder, H. (1904). The Geographic Pivot of History. *Geographical Journal*, 23(4), 421 – 441.

Mandel, R. (1980). Roots of Modern Interstate Border Disputes. *Journal of Conflict Resolution*, 24, 427 – 454.

Marshall, M. G., Gurr, T. R., & Jaggers, K. (2017). Political Regime Characteristics and Transitions, 1800 – 2016 [Data file]. Retrieved from http://www.systemicpeace.org/inscrdata.html.

Menard, S. (2002). Quantitative Applications in the Social Sciences: Applied logistic regression analysis. Thousand Oaks, CA: SAGE Publications.

Midlarsky, M. (1975). *On War*. New York: Free Press.

Moran, P. A. P. (1948). The Interpretation of Statistical Map. *Journal of the Royal Statistical Society Series B*, 10, 245 – 251.

Morgenthau, H. J. (1967). *Politics Among Nations: The Struggle for Power and Peace*. New York: Knopf.

Most, B. & Starr, H. (1980). Diffusion, Reinforcement, Geopolitics, and the Spread of War. *American Political Science Review*, 74(4), 932 – 946.

Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2015). Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced. *Political Analysis*, 24, 87 – 103.

Murdoch, J. & Sandler, T. (2002). Economic Growth, Civil Wars, and Spatial Spillovers. *Journal of Conflict Resolution*, 46(1), 91 – 110.

Murshed, S. M. & Gates, S. (2005). Spatial-Horizontal Inequality and the Moaist Insurgency in Nepal. *Review of Development Economics*, 91, 121 – 134.

National Consortium for the study of Terrorism and Reponses to Terrorism (START). (2019). Global Terrorism Database [Data file]. Retrieved from http://www.start.umd.edu/gtd.

Nelder, J. A. & Wedderburn. (1972). Generalized Linear Regression. *Journal of the Royal Statistical Society*, 135(3), 370 – 384.

Nemeth, S. C., Mauslein, J. A., & Stapley, C. (2014). The Primacy of the Local: Identify Terrorist Hot Spots Using Geographic Information Systems. *The Journal of Politics*, 76(2), 304 – 317.

Nordquist, K. (1986). *The Settlement of Border Conflicts: A Theoretical Model with Empirical Illustrations*. Presented at IPRA conference: Sussex, England.

Nye, Jr, J. S. (2003). Understanding International Conflicts: An Introduction to Theory and History. New York: Longman.

O'Brien, S. P. (2010). Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *International Studies Review*, 12(1), 87 – 105.

O'Loughlin, J. (1986). Spatial Models of International Conflicts: Extending Current Theories of War Behavior. *Annals, Association of American Geographers*, 76(1), 63 – 80.

O'Loughlin, J. & Anselin, L. (1991). Bringing Geography Back to the Study of International Relations: Spatial Dependence and Regional Context in Africa, 1966-1978. *International Interactions*, 17(1), 29 – 61.

O'Loughlin, J. & Witmer, F. (2009). The Localized Geographies of Violence in the North Caucasus of Russia, 1999-2007. *Annals of the Association of American Geographers*, 101(1), 178 – 201.

Openshaw, S., Charlton, M., Craft, A.W. & Birth, J.M. (1988). Investigation of Leukemia clusters by the use of a geographical analysis machine, Lancet, I, 272-273.

Ord, J. & Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis*, 27, 286 – 306.

Paelinck, J. H. P. & Klaassen, L. H. (1979). *Spatial Econometrics*. Farnborough: Saxon House.

Perry, C. (2013). Machine Learning and Conflict prediction: A Use Case. *Stability: International Journal of Security & Development*, 2, 1 – 18.

Prescot, J. R. V. (1965). *The Geography of Frontiers and Boundaries*. Chicago: Aldine.

Raleigh, C. & Hegre, H. (2009). Population Size, Concentration, and Civil War: A Geographically Disaggregated Analysis. *Political Geography*, 28(4), 224 – 238.

Raleigh, C., Linke, A., Hegre, H., & Karlsen, J. (2010). Introducing ACLED: Armed Conflict Location and Event Data. *Journal of Peace Research*, 47(5), 651 – 660.

Raleigh, C., Witmer, F., & O'Loughlin, J. (2010). The Spatial Analysis of War, in Robert Denemark, ed., *The International Studies Encyclopedia*, Vol. X. Oxford, UK: Wiley-Blackwell, 6534 – 6553.

Richardson, L. (1960). *Statistics of Deadly Quarrels*. Pittsburgh: Boxwood.

Rogerson, P. A. (2015). *Statistical Methods for Geography: A Student's Guide*. Los Angeles: Sage.

Rustad, S.C., Buhaug, H., & Falch, A., & Gates, S. (2011). All Conflict is Local: Modeling Sub-National Variation in Civil Conflict Risk. *Conflict Management and Peace Science*, 28(1), 15 – 40.

Salehyan, I. (2009). *Rebels without Borders: Transnational Insurgencies in World Politics*. Ithaca, NY: Cornell University Press.

Salehyan, I., & Gleditsch, K. S. (2006). Refugees and the Spread of Civil War. *International Organization*, 60(2), 335 – 366.

Salehyan, I., Hendrix, C. S., Hamner, J. Case, C., Linebarger, C., Stull, E., & Williams, J. (2012). Social Conflict in Africa: A New Database. International Interactions, 4, 503 – 511.

Sambanis, N. (2001). Do Ethnic and Nonethnic Civil Wars Have the Same Causes?: A Theoretical and Empirical Inquiry. *Journal of Conflict Resolution*, 45(3), 259 – 282.

Sambanis, N. (2004). What Is Civil War? Conceptual and Empirical Complexities. *Journal of Conflict Resolution, 48*(6), 814 – 858.

Schrodt, P. A. (1988). Artificial Intelligence and the Study of International Politics. *American Sociologist*, 19(1), 71 – 85.

Schrodt, P. A. (1991a). Pattern Recognition of International Event Sequences: A Machine Learning Approach. *Artificial Intelligence and International Politics*, 169 – 193.

Schrodt, P. A. (1991b). Prediction of Interstate Conflict Outcomes Using a Neural Network. *Social Science Computer Review*, 9(3), 359 – 380.

Schrodt, P. A. (1999). Early Warning of Conflict in Southern Lebanon using Hidden Markov M Models. In Starr, H. eds., *The Understand and Management of Global Violence: New Approaches to Theory and Research of Protracted Conflict*, New York City: St. Martin's Press.

Schrodt, P. A. (2000). Pattern Recognition of International Crises Using Hidden Markov Models. In Richards, D. eds. *Political Complexity: Nonlinear Models of Politics*, Ann Arbor: University of Michigan Press.

Schrodt, P. A. (2006). Forecasting Conflict in the Balkans using Hidden Markov Models. In Trappl, R. eds. *Programming for Peace: Computer-Aided Methods for International Conflict Resolution and Prevention*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Schrodt, P. A. (2014). Seven Deadly Sins of Contemporary Quantitative Political Analysis. Journal of Peace Research, *Journal of Peace Research*, 51(2), 287 – 300.

Schrodt, P. A., Davis, S. G., & Weddle, J. L. (1994). Political Science: KEDS—A Program for the Machine Coding of Event Data. *Social Science Computer Review*, 12(4), 561 – 587.

Schrodt, P. A. & Gerner, D. J. (2000). Cluster-based Early Warning Indicators for Political Change in the Contemporary Levant. *American Political Science Review*, 94(4), 803 – 817.

Schrodt, P. A., Yonamine, J., Bagozzi, B. E. (2013). Data-based Computational Approaches to Forecasting Political Violence. In Subrahmanian V. (eds) *Handbook of Computational Approaches to Counterterrorism* (129 – 162). New York: Springer.

Schutte, S. (2016). Regions at Risk: Predicting Conflict Zones in African Insurgencies. *Political Science Research and Methods*, 5, 1 – 19.

Schutte, S. & Weidmann, N. B. (2011). Diffusion Patterns of Violence in Civil Wars. *Political Geography*, 30(3), 143 – 152.

Senese, P. D. (2005). Territory, Contiguity, and International Conflict: Assessing a New Joint Explanation. *American Journal of Political Science*, 49(4), 769 – 779.

Shearer, R. (2007). Forecasting Israeli – Palestinian Conflict with Hidden Markov Models. *Military Operations Research*, 12(4), 5 – 15.

Shuttle Radar Topography Mission (SRTM). (2014). SRTM Africa Images [Data file]. Retrieved from www2.jpl.nasa.org/srtm/Africa_radar_images.htm.

Siroky, D. S. (2009). Navigating Random Forests and Related Advances in Algorithmic Modeling. *Statistics Surveys*, 3, 147 – 163.

Siverson, R. & Starr, H. (1991). *The Diffusion of War: A Study of Opportunity and Willingness*. Ann Arbor, MI: University of Michigan Press.

Sprout, H. & Sprout, M. (1965). *The Ecological Perspective on Human Affairs*. Princeton: Princeton University Press.

Sorokin, P. A. (1957). *Social and Cultural Dynamics*. New York: American Book Company.

Spykman, N. (1938). Geography and Foreign Policy, I. *American Political Science Review*, 32, 28 – 50.

Starr, H. (1978). Opportunity and Willingness as Ordering Concepts in the Study of War. *International Interactions*, 4, 363 – 387.

Starr, H. (2005). Territory, Proximity, and Spatiality: The Geography of International Conflict. *International Studies Review*, 7(3), 387 – 406.

Starr, H. & Most, B. (1976). The Substance and Study of Borders in International Relations. *Journal of Conflict Resolution*, 20, 581 – 620.

Starr, H. & Most, B. (1983). Contagion and Border Effects on Contemporary African Conflict. *Comparative Political Studies*, 16, 92 – 117.

Starr, H. & Most, B. (1985). The Forms and Processes of War Diffusion: Research Update on Contagion in African Conflict. *Comparative Political Studies*, 18, 206 – 227.

Sundberg, R. & Melander, E. (2013). Introducing the UCDP Georeferenced Event Dataset. *Journal of Peace Research*, 50(4), 523 – 532.

Theisen, O. M. (2012). Climate Clashes? Weather Variability, Land Pressure, and Organized Violence in Kenya, 1989 – 2004. *Journal of Peace Research*, 49(1), 81 – 96.

Theisen, O. M., Holtermann, H., Buhaug, H. (2012). Climate Wars?: Assessing the Claim That Drought Breeds Conflict. *International Security*, 36(3), 79 – 106.

Tobler, W. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46, 234 – 240.

United Nations Office for the Coordination of Humanitarian Affairs (OCHA). (2019). *Administrative Boundaries* [Data file]. Retrieved from https://data.humdata.org.

Ward, M. D. & Gleditsch, K. (2002). Location, Location, Location: An MCMC Approach to Modeling the Spatial Context of War and Peace. *Political Analysis*, 10, 244 – 260.

Ward, M. D., Greenhill, B. D., Bakke, K. M. (2010). The Perils of Policy by p-value: Predicting Civil Conflict. *Journal of Peace Research*, 47, 363 – 375.

Ward, M. D., Metternich, N. W., Dorff, C. L., Gallop, M., Hollenbach, F. M., Schultz, A., & Weschle, S. (2013). Learning from the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction. *International Studies Review*, 15(4), 473 – 490.

Webb, K. (2007). The Continued Importance of Geographic Distance and Boulding's Loss of Strength Gradient. Comparative Strategy, 4, 295 – 310.

Weezel, S. V. (2017). Predicting Conflict Events in Africa at Subnational Level.

Weidmann, N. B. (2009). Geography as Motivation and Opportunity: Group Concentration and Ethnic Conflict. *Journal of Conflict Resolution*, 53 (4), 526 – 543.

Weidmann, N. B., Rød, J. K., & Cederman, L. (2010). Representing Ethnic Groups in Space: A New Dataset. Journal of Peace Research, 47(4), 491 – 499.

Weidmann, N. B. & Ward, M. D. (2010). Predicting Conflict in Space and Time. Journal of Conflict Resolution, 54(6), 883 – 901.

Weschle, S. (2013). Learning From the Past and Stepping into the Future: Toward a New Generation of Conflict Prediction. *International Studies Review*, 15(4), 473 – 490.

Witmer, F. DW., Linke, A. M., O'Loughlin, J. Gettelman, A. & Laing, A. (2017). Subnational Violent Conflict Forecasts for Sub-Saharan Africa, 2015–65, Using Climate-Sensitive Models. Journal of Peace Research, 54(2), 175 – 192.

Wolff, S. (2006). *Ethnic Conflict: A Global Perspective*. Oxford: Oxford University Press.

World GeoDatasets. (2015). *World Language Mapping System, Version 16* [Dataset]. Retrieved from http://www.worldgeodatasets.com/language/.

Wright, Q. (1942). *A Study of War*. Chicago: University of Chicago Press.

Yonamine, J. E. (2013). Predicting Future Levels of Violence in Afghanistan Districts Using GDELT. *GDELT Working Paper*, UT – Dallas.